PROKARIÓTA GENOMOK ÖSSZEHASONLÍTÓ ANALÍZISE BIOINFORMATIKAI MÓDSZEREKKEL

Doktori (Ph.D.) értekezés

Kassainé Jáger Edit Andrea

Biológia Doktori Iskola

Vezetője: Dr. Erdei Anna egyetemi tanár, az MTA l. tagja

Elméleti- és evolúcióbiológia Doktori Program

Programvezető: Dr. Szathmáry Eörs, D.Sc., egyetemi tanár, az MTA l. tagja

Témavezető: Dr. Vida Gábor, D.Sc., az MTA rendes tagja

ELTE TTK Genetikai Tanszék

Budapest

2008

Tartalomjegyzék

Az értekezés alapjául szolgáló közlemények

1. Ortutay, C., Gáspári, Z., Tóth, G., Jáger, E., Vida, G., Orosz, L. and Vellai, T. (2003) Speciation in Chlamydia: genomewide phylogenetic analyses identified a reliable set of acquired genes. *J Mol Evol.*, **57**, 672-680.

2. Kassai-Jáger, E., Ortutay, C., Tóth, G., Vellai, T. and Gáspári, Z. (2008) Distribution and evolution of short tandem repeats in closely related bacterial genomes. *Gene*, **410**, 18-25.

Ábrák jegyzéke

Táblázatok jegyzéke

Rövidítések jegyzéke

bp	bázispár
BLAST	lokális illesztéseken alapuló, szekvenciahasonlóságot kereső program (*B*asic *L*ocal *A*lignment *S*earch *T*ool)
CAI	kodon adaptációs index (*C*odon *A*daptation *I*ndex)
CDS	kódoló szekvencia (*Cod*ing *S*equence)
CDT	egyesített távolság (*C*umulative *D*is*T*ance)
COGs Database	az ortológ fehérjecsoportok klaszterjeit tartalmazó adatbázis (*C*lusters of *O*rthologous *G*roups Database, ftp://ftp.ncbi.nih.gov/pub/COG)
CLUSTALW	többszörös szekvencia illesztéseket (Multiple Sequence Alignments) készítő program
E érték	várható érték, (*E*xpected Value)
E. coli	*Escherichia coli*
EHEC	Enterohemorrhágiás *Escherichia coli* (*E*ntero*h*emorrhagic *Escherichia coli*)
FASTA	DNS és fehérje szekvencia illesztő programcsomag (DNA and Protein sequence alignment software package)
GC - tartalom	guanin - citozin tartalom
HGT	horizontális géntranszfer (*H*orizontal *G*ene *T*ransfer)
KEGG	a kiotói egyetem Gén és Genom Enciklopédiája (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg)
Mbp	megabázispár (10^6 bázispár)
MySQL	több felhasználós, többszálú, SQL-alapú relációs adatbázis-kezelő szerver, az egyik legelterjedtebb adatbázis kezelő (http://www.mysql.com)
NCBI	Nemzeti Biotechnológiai Információs Központ, Bethesda, Maryland, USA (*N*ational *C*enter for *B*iotechnology *I*nformation, Bethesda, Maryland, USA, http://ncbi.nih.gov/)
NIIF	*N*emzeti *I*nformációs *I*nfrastruktúra *F*ejlesztési *I*ntézet
NIH	az amerikai Nemzeti Egészségügyi Intézet (*N*ational *I*nstitutes of *H*ealth)
nt	nukleotid

ORF	nyílt leolvasási keret (*O*pen *R*eading *F*rame)
OTU	kezelendő Taxonómiai Egység (*O*perational *T*axonomic *U*nit)
`PAUP*`	Filogenetikai Analízis Parszimónium Használatával programcsomag (*P*hylogenetic *A*nalysis *U*sing *P*arsimony)
Pmps	polimorfikus membránfehérjék (*P*olymorphic *M*embrane *P*roteins)
`PRIDE`	Cα atomok közötti távolságok eloszlásán alapuló térszerkezet-összehasonlító módszer (*P*robability of *IDE*ntity)
SSRs	egyszerű szekvenciális ismétlődések, mikroszatelliták (*S*imple *S*equence *R*epeats)
`TREE-PUZZLE`	Maximum likelihood analízist végrehajtó program (Maximum likelihood analysis for nucleotide, amino acid, and two-state data)
`TRF`	tandem ismétlődéseket azonosító program (`Tandem Repeats Finder`)
UPEC	uropatogén *Escherichia coli* (*Urop*athogenic *E*scherichia *c*oli)
VNTR	változó számú tandem ismétlődések (*V*ariable *N*umber of *T*andem *R*epeats)
WHO	Egészségügyi Világszervezet (*W*orld *H*ealth *O*rganization)

1. Bevezetés

1.1. A horizontális géntranszfer

Az egyre több prokarióta genomi szekvencia megismerésével az utóbbi években megnyílt az út ezen szervezetek evolúciójának részletes és átfogó kutatásához.

A horizontális géntranszfer (*H*orizontal *G*ene *T*ransfer, HGT), más néven laterális géntranszfer egy olyan esemény, amikor a genetikai információ átadása nem szülő- és utódszervezetek között zajlik, mint a vertikális géntranszfer esetében. A HGT gyakran fordul elő a bakteriális evolúció során és jelentősen hozzájárul a baktériumok diverzitásához és adaptációjához (Ochman 2000). Mindazonáltal nem könnyű meghatározni az adott bakteriális genomon belül a HGT által érintett gének hányadát, és nehéz feladat annak megállapítása is, hogy milyen mértékben vesz részt ez a mechanizmus a genomszerkezet változtatásában. Azokat a géneket, melyek az evolúciós időskála szerint régen érkeztek a genomba általában nehezebben lehet detektálni, mint a közelmúltban „transzferáltakat", mivel más folyamatok (pl. génvesztés vagy diverzifikáció) is befolyásolhatják az analízist. Napjainkban három különböző módszert alkalmaznak a HGT kimutatására. A legpontosabb, de nagy számítógépes kapacitást igénylő módszer a kiválasztott gének részletes filogenetikai analízise (Sicheritz-Ponten *és* Andersson 2001). Ez a típusú analízis számítási igénye miatt jelenleg nem alkalmas teljes genomok horizontálisan transzferált génekre vonatkozó vizsgálatára. A másik kettő, sokkal inkább és széleskörűbben alkalmazott módszer, a szervezet összes génjével végzett szisztematikus hasonlósági keresés egy kellően nagy adatbázissal szemben, valamint a kódoló régiók nukleotid összetételének és/vagy a kodonhasználat vizsgálatán alapul.

1.2. A mikroszatelliták

A bakteriális genomok összevetése/összehasonlítása nem lehet teljes a mikroszatelliták részletes jellemzése nélkül. A mikroszatelliták vagy más néven egyszerű szekvenciális ismétlődések (*S*imple *S*equence *R*epeats, SSR-ek) gyakorlati és elméleti jelentőségét eukariótákban többen leírták (Ellegren 2004; Kashi *és* King 2006; Tóth *és mtsi.* 2000). Genetikai markerként széleskörűen alkalmazzák őket, ezen túl a genomevolúcióban betöltött szerepük miatt is fontos vizsgálatuk. Született néhány a prokarióta mikroszatellita evolúcióval kapcsolatos munka (Eckert *és* Yan 2000; Metzgar *és mtsi.* 2001; Schlotterer *és mtsi.* 2006; Mrazek *és mtsi.* 2007), sőt még egy internetes szolgáltatás is igénybe vehető (http://insilico.ehu.es/microsatellites/info.html), ezentúl a változó számú tandem ismétlődés (*V*ariable *N*umber of *T*andem *R*epeats, VNTR) polimorfizmust is elemezték az *Escherichia coli* O157:H7 törzsben (Noller *és mtsi.* 2003), és megállapítást nyert, hogy az SSR-ek kétségtelenül hozzájárulnak a bakteriális genom polimorfizmushoz (Lindstedt 2005). Mindezek ellenére a mikroszatellita-eloszlás baktériumokban kevésbé tanulmányozott, mint az eukariótákban, főként annak köszönhetően, hogy az SSR-ek viszonylag kisebb gyakorisággal fordulnak elő a prokarióta genomokban (Ellegren 2004). A több mint 350[a] prokarióta genomi szekvenciához való hozzáférési lehetőség az SSR analízisre vonatkozó új technikák megjelenésével lehetőség nyílt az SSR-ek részletes és átfogó kutatására prokariótákban is.

A közeli rokon bakteriális genomokban megfigyelhető mikroszatellita-eloszlás összehasonlításával közvetlen becslést lehet végezni ezen baktériumok evolúciójára vonatkozóan. Az SSR-eket tartalmazó géneket vizsgálva a rokon törzsekben bepillantást nyerhetünk abba a folyamatba, ahogyan az egyszerű ismétlődések formálják a bakteriális fehérjék szerkezetét.

[a]2004-ben, mikroszatellita vizsgálataink megkezdésekor érvényes adat, a dolgozat írásának időpontjában (2008. április) 682 teljes mikrobiális genomszekvencia található az Entrez Genome adatbázisban (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1).

Genomi szinten az egyik lehetséges megközelítés a tökéletes és a nem tökéletes ismétlődések független felismerése és összehasonlítása. Technikai szempontból, mivel nincs egységesen elfogadott mikroszatellita definíció (Ellegren 2004), a különböző tanulmányokban azonosított SSR-ek összehasonlítása nem egyszerű feladat. Áthidaló megoldásként, új megközelítésként, a kétféle kimutatás konszenzusát alkalmazhatjuk (Gáspári *és mtsi*. 2007), a tökéletes ismétlődéseket és a nem tökéletes ismétlődéseket külön, függetlenül azonosítjuk, majd ezt követően összehasonlítjuk. Ezzel a megközelítéssel információt nyerhetünk az ismétlődések történetéről, ha feltételezzük, hogy a tökéletes ismétlődés „mag"-ot (core) tartalmazó nem tökéletes ismétlődések többsége egy hosszabb tökéletes ismétlődés szakasz maradványa. Ily módon a nem tökéletes ismétlődések és ezek tökéletes ismétlődésű „mag" szegmensének hosszbeli eloszlásában tapasztalható különbségek/eltérések az ismétlődések kiterjedésével, kiterjesztésével, ill. lebomlásával, pusztulásával, kapcsolatos evolúciós eseményeket jelezhetnek (Gáspári *és mtsi*. 2007).

A kapcsolódó genomok párhuzamos vizsgálatát és a standardizált SSR klasszifikációt alkalmazó többszörös SSR detektáló módszereket (Jurka *és* Pethiyagoda 1995; Tóth *és mtsi*. 2000) együtt használva várhatóan biológiai szempontból releváns képet nyerhetünk az SSR-ek jelentőségéről a vizsgált bakteriális törzsekben. Fontos kérdés az SSR-ek evolúciójuk során mutatott mutabilitása (Kashi *és* King 2006), amit genomi szinten a tökéletes és a nem tökéletes ismétlődések összehasonlításával lehet tanulmányozni.

2. Irodalmi áttekintés

2.1. A *Chlamydiales* rend

A *Chlamydiák* (*Chlamydiae*) a *Chlamydiales* bakteriális rendbe tartozó obligát intracelluláris baktériumok. Sok *Chlamydia* él együtt tünetmentes állapotban meghatározott gerincesek testében vagy amőbákban, széles körben elterjedt a nézet, miszerint ezek a gazdák természetes tárolóhelyet (reservoir-t) biztosítanak ezen fajok számára (Everett 2000a). A *Chlamydiales* rend nagyon közeli rokonokból álló monofiletikus csoportot alkot, amely filogenetikailag elkülönül a többi bakteriális taxontól (Kalman *és mtsi.* 1999; Read *és mtsi.* 2000, 2003; Shirai *és mtsi.* 2000; Stephens *és mtsi.* 2000). A csoport tagjai közül a három *Chlamydophyla pneumoniae* törzs (a CWL029, az AR39 és a J138*)* és a két *Chlamydia* törzs (a *Chlamydia trachomatis* D/UW-3/CX és az egérpatogén *Chlamydia muridarum* Nigg) viszonylag kis genommérettel rendelkeznek, ami a reduktív genomevolúció egyik példája. Ezen túl az összehasonlító genomika viszonylag alacsony szintű DNS szintű hasonlóságot mutatott ki a *C. trachomatis* és a *C. pneumoniae* törzsek között. Ezen genom tulajdonságok alapján választottuk ki a fenti 5 *Chlamydiá*t egy genom szintű összehasonlító analízishez.

2.1.1. A *Chlamydiák*[a] által okozott betegségek

A *Chlamydiák* rendjébe klinikai szempontból fontos, obligát intracelluláris parazita (állat- és humánpatogén) baktériumok és eukarióta gazdák endoszimbiontái tartoznak. Az általuk okozott fertőzések gyakran súlyos utólagos következményekkel járnak. A *Chlamydiák* sok vadon élő és háziasított állatban is gyakoriak, potenciális komoly zoonózis veszélyt jelentenek (Everett 2000b). A *Chlamydia trachomatis* és a *Chlamydophila pneumoniae* humánpatogének, de a madárpatogén *Chlamydophila psittaci* is okozhat súlyos tüdőgyulladást, papagájkórt, ha a kórokozó emberekre is átterjed.

[a]Jelen értekezésben a fajneveket, törzsneveket illetően az érvényes „új" taxonómiát követem (Bush *és* Everett 2001), de a „*Chlamydia"* egyesítő kifejezést is használom, amikor a *Chlamydia* és *Chlamydophila* genusokról együttesen van szó (lásd 19. oldal).

A *Chlamydia trachomatis* szem- és nemiszerv fertőzéseket okoz. A *Chlamydia trachomatis* fenotípusos jellemzése történhet a baktérium anyagcsere vizsgálatával (biokémiai, biovar), illetve a sejtantigének meghatározásával (szerológiai, serovar).

Szerológiai alapon három *Chlamydia trachomatis* csoportot különítenek el. Az első két csoport az A-C serovar és a D-K serovar alkotja a *trachoma biovar* csoportot, a harmadik csoport az L1-L3 serovar pedig az LGV biovarral azonos. Az A-C serovar a fejlődő országokban endemikus[b] egyiptomi szemgyulladást (trachoma), okoz, mely kezeletlenül vaksághoz vezethet. A D-K serovarba tartozó szexuális úton terjedő kórokozók különféle húgy-ivarszervi betegségeket (húgycsőgyulladás, méhnyakgyulladás és petevezetékgyulladás) okoznak. A fertőzés gyakran tünetmentes és meddőséget okozhat, valamint növeli a méhen kívüli terhesség kockázatát is. Az L1-L3 serovarba tartozó baktériumok is egy szexuális úton terjedő betegséget, a Nicolas-Durand-Favre-kórt (*L*ymphogranuloma *V*enerum, LGV) okoznak, amely a lágyéki nyirokcsomók gyulladását idézi elő. A LGV nagyon komoly fertőzés, mert könnyen szétterjed a nyirokkeringésben és rendszeressé (szisztémássá) válhat. A *Chlamydophila pneumoniae* akut és krónikus légzőszervi betegségeket okoz, kapcsolatba hozható az asztmával és az érelmeszedés egy fajtájával (atherosclerosis) is, melyet az artériák belső rétegében zsírszerű anyaglerakódások okoznak (Kalman *és mtsi*. 1999; Read *és mtsi*. 2000; Vandahl 2004).

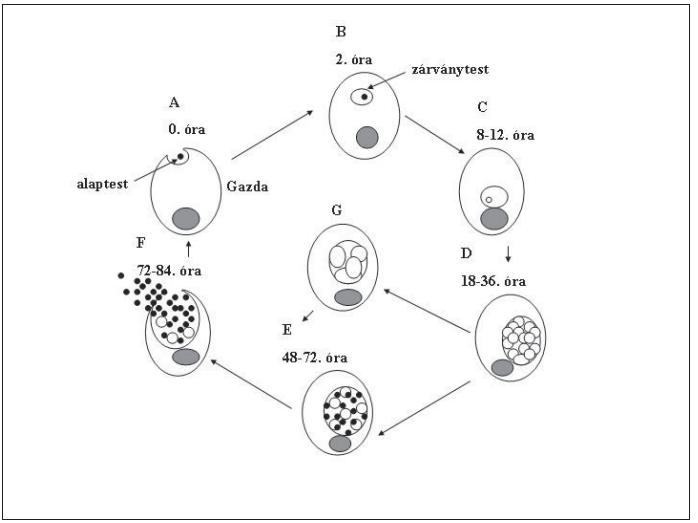
2.1.2. A *Chlamydiák* életciklusa

A *Chlamydiák* Gram negatív, obligát intracelluláris paraziták, kétfázisú, egyedül a *Chlamydiák*ra jellemző fejlődési ciklusuk (1.ábra) során a baktériumok két alakja váltakozik, egy fertőzőképes alaptest (elementary body, EB) és egy nem fertőzőképes sejtbeli replikatív alak a hálózatos test (reticulate body, RB). Az alaptestek kicsi, kb. 300 nm átmérőjű testecskék, melyeket általában metabolikusan inaktívként szoktak jellemezni, DNS-ük kompakt szerkezetű, hisztonszerű fehérjéik vannak. A sejten kívüli (extracelluláris) élethez alkalmazkodtak, ozmotikus stabilitást biztosító külső membránjukban erős diszulfid keresztkötések találhatók. A hálózatos testek kb. 1 μm átmérőjű, külső membránjuk permeábilis a gazdasejt tápanyagai számára átengedi a gazdasejt tápanyagait, DNS-ük - hasonlóan más baktériumokéhoz- nincs bepakolódva.

[b]helyi járvány, népbetegség

A fertőző hálózatos test megközelíti a fogékony gazdasejtet, ami aztán fagocitálja. A pontos mechanizmust nem ismerik, de a felvételt valamiképpen a baktérium indukálja. A fagoszóma belsejében, amit zárványnak (inclusion) neveznek, az alaptest kettéosztódással hálózatos testté fejlődik. Ez a folyamat magában foglalja a DNS kicsomagolását és a külső membrán diszulfidhídjainak redukálódását, csökkentését, azt azonban nem lehet tudni, hogy mi váltja ki ezeket az eseményeket. Többszöri osztódások után a hálózatos testek elkezdenek alaptestekké alakulni, DNS-ük becsomagolódik, és a későbbi külső membrán is megszintetizálódik Végül a fertőzőképes alaptestek új generációja kiszabadul a gazdasejt felbomlásával. A baktériumok a zárványban maradnak a teljes intracelluláris fázis során, ami a *Chlamydia pneumoniae* sejtkultúrákban 72-96 óráig tart. A *Chlamydiák* a zárvány membránjának felépítéséhez a gazdasejt lipidjeit használják fel, amit aztán az ún. zárvány membránfehérjék (*inc*lusion membrane protein*s*, incs) beépítésével módosítanak. A *Chlamydia pneumoniae* fejlődési ciklusa felfüggeszthető a gazdasejt interferon-gamma indukált katabolizmusával. A triptofánéhezés nem-termelő fertőzéshez vezet, melyben megnövekedett, aberrált hálózatos testek keletkeznek. Ezek nem osztódnak és nem alakulnak át érett alaptestekké, de a normális fejlődési ciklust vissza lehet állítani. A *Chalmydia trachomatis* is képes kitartó képletté alakulni, a citokineken kívül a limitált tápanyag ellátottságról és az antibiotikumkezelésről (amely elpusztítja a kórokozót) mutatták ki, hogy kiválthatja ezt az állapotot (Vandahl 2004; U.S. CDC[c] 2007).

[c]CDC: Betegség Ellenőrzési és Megelőzési Központok (*Ce*nters for *D*isease *C*ontrol and Prevention), az amerikai Egészségügyi és Humán Szolgálati Hivatal (U.S. Department of Health and Human Services HHS.gov) egyik szervezete



1. ábra. A *Chlamydiák* életciklusa

Az ábrán a *C. pneumoniae* sejttenyészetben a fertőzés után eltelt időt órákban kifejezve adtuk meg. **A:** a fertőzőképes alaptest a gazdasejt felszínéhez kötődik és a gazdasejtbe jut (endocitózis); **B:** a *Chlamydia* átalakítja a fagoszómát, zárványtestté alakul, hogy elkerülje az endocitotikus útvonalat; **C:** az alaptest a metabolikusan aktív hálózatos testté fejlődik; **D:** a hálózatos testek kettéosztódnak és a zárványtest megnő a gazdasejtből származó lipidek bekebelezésével **E:** többször osztódás után a hálózatos testek újraszerveződnek és átalakulnak alaptestekké; **F:** végül a fertőzőképes alaptestek új generációja a gazdasejt lízisével kiszabadul; **G:** az alacsony tápanyagellátottság, az interferon-γ közvetített triptofán éhezés vagy más stresszt jelentő körülmény egy olyan abnormális, állandó/folyamatos állapot létrejöttét idézheti elő, amely nem osztódó hálózatos testek jelenlétével jellemezhető. Ezek a hálózatos testek újra aktiválhatók és visszatérhetnek a fejlődési ciklusba, amikor a feltételek megfelelők a növekedéshez. (Vandahl 2004 nyomán, ábra átrajzolva).

2.1.3. A *Chlamydiák* taxonómiája[d]

Amikor 1980-ban az IJSB publikálta a jóváhagyott baktérium listát, a *C. trachomatis*t és a *C. psittaci*t tartották meg, a korábbi *Chlamydia* nevek 90%-át elvetették. Változás 1988 után kezdődött, amikor a DNS-DNS reasszociációt alkalmazták a *Chlamydia* törzsek megkülönböztetésére (Everett 2000b). A *C. psittaci*n belüli meghökkentő diverzitás nyilvánvalóvá vált (Cox *és mtsi.* 1988; Grayston *és mtsi.* 1989). A legtöbb vizsgált *C. psittaci* 70%-nál nagyobb DNS hibridizációs hasonlóságot mutatott egymással, egy olyan megkülönböztető jegyet, amelyet Schleifer *és* Stackebrandt (1983) alkottak meg a baktérium fajok elkülönítésére. Ezen kritériumok alapján a *C. psittaci* (ami akkor a *C. pecorum*ot és a *C. pneumoniae* törzseket is magában foglalta) világosan látszott, hogy legalább 6 különböző fajt tartalmaz (Everett 2000a). A 21. századi *Chlamydia* taxonómia felé vezető első lépés az új információ felhalmozása volt, elsődlegesen a riboszómális szekvencia adatoké. A riboszóma szekvencia analízisek alapján a *Chlamydia* és a *Chlamydophila* ág divergálódásának elméletét dolgozták ki.

Napjainkban a mikrobiológusok világos kritériummal rendelkeznek a nemzetség azonosítására, a közel rokon baktérium nemzetségekben a 16S rRNS szekvenciának 95%-ban azonosnak kell lennie. A taxonómiai döntést numerikus és evolúciós kritériumok alapján hozzák meg. A taxonómiában történt változások azért játszódtak le, mert új információkat szereztünk. A szisztematikusok egyetértenek abban, hogy a fenotípusos, genetikai és a filogenetikai analíziseket egyesíteni kell amennyire lehet, hogy a taxonómia alapját képezhessék. Ezen erőteljes/hatékony módszerek egységesítése forradalmasította a taxonómiát általában, és specifikusan a *Chlamydia* taxonómiát is (Everett 2000b).

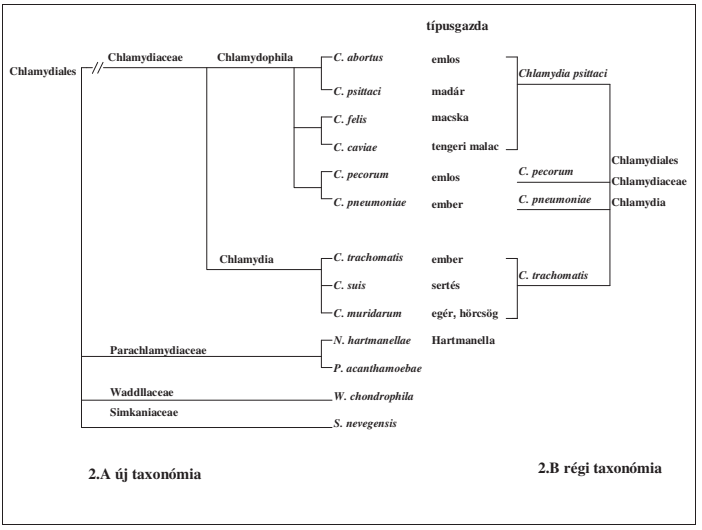
[d]Taxonómia: a biológiai sokféleséget (biodiverzitást) a jelenségek szintjén vizsgáló, tapasztalati, elemző, részekre bontó, leíró, azonosító, névadó, összehasonlító, elrendező tudomány. Döntően maguknak az élőlényeknek a közvetlen vizsgálatán, jelelmzésén és ez alapján való csoportosításán alapszik.

A Bakterológiai Törvénykönyv szerint egy baktériumot csak úgy lehet elnevezni, ha publikálják az Amerikai Mikrobiológiai Társaság (*A*merican *S*ociety of *M*icrobiology, ASM) által kiadott IJSB (a Baktériumok rendszertanának nemzetközi folyóirata, *I*nternational *J*ournal of *S*ystematic *B*acteriology), a későbbi IJSEM (a mikrobiológia rendszertani és evolúciós kérdéseinek nemzetközi folyóirata, *I*nternational *J*ournal of *S*ystematic and *E*volutionary *M*icrobiology) tudományos folyóiratban. Ez a baktérium klasszifikáció a plusz/mínusz adatokkal dolgozó numerikus taxonómián alapult. A numerikus taxonómiát 1956 óta alkalmazzák baktériumokra. Csak azokat a karaktereket vették számításba, amelyek az adott szervezet genotípusának direkt vagy indirekt eredményei, pl. morfológiai, biokémiai, élettani jellegzetességek, tápanyagszükséglet, antigén összetétel, fágszenzitivitás. Genetikai adatok, tesztek híján a *Chlamydiales* rendbe csak két faj került (Everett 2000a).

1999-ben egy olyan filogenetikai[e] kapcsolatokon alapuló új *Chlamydia* taxonómia bevezetését javasolták, több új nemzetség és faj bevezetésével (2. ábra), amely 95%-nál nagyobb 16S rRNS azonosságot követel meg nemzetségen belül. Az új családok, nemzetségek és fajok elkülönítését támogatta az összes elfogadott bakteriális szisztematikai standard (Stackebrandt *és* Goebel 1994; Palys *és mtsi*. 1997; Ludwig *és mtsi.* 1998; Everett 2000a). A javasolt rendszer a *C. trachomatis*t a *Chlamydia* nemzetségbe tette és a korábbi *C. trachomatis*t 3 új fajra osztotta. A megmaradó *Chlamydia* fajokat egy új, a *Chlamydophila* nemzetségbe sorolta és a *C. psittaci*t több különböző fajra bontották (Everett *és mtsi*. 1999; Bush *és* Everett 2001; Vandahl 2004). A *C. suis* törzseket korábban a *C. trachomatis*ba sorolták az *ompA* DNS szekvencia hasonlóság miatt, a korábbi *Chlamydia pecorum*ot *Chlamydophila pecorum*ra nevezték át. A macskaféléket megbetegítő *Chlamydophila felis* főként kötőhártyagyulladást, hurutos orrnyálkahártya-gyulladást (nátha) és légzőszervi problémákat okoz. A *Chlamydophila caviae* szembetegséggel rendelkező tengeri malacokból nyerhető ki, a *Chlamydophila abortus*t évekig *C. psittaci*ként jellemezték (Everett 2000a).

Az újonnan leírt *Chlamydia* fajok, a *Simkania* és a *Parachlamydia* is humán légzőszervi fertőzésekkel állnak kapcsolatban, míg a *Waddlia* szarvasmarha vetélésekkel hozható kapcsolatba (Kahane *és mtsi.* 1998; Horn *és mtsi*. 2000; Corsaro *és* Venditti 2004; Griffith *és mtsi*. 2006). Ezek a revíziók fontos változásokat tükröznek a *Chlamydia* diverzitás érzékelésével kapcsolatban, összhangban az új állati izolátumok (*Waddliae* és *Simkaniae*), és a még meglepőbb szabadon élő amőbákat megfertőző *Chlamydia*-rokon endoszimbionták vagy „környezeti *Chlamydiae*" (azaz *Parachlamydiae*) felfedezésével (Corsaro *és mtsi.* 2003; Horn *és* Wagner 2001; Fritsche *és mtsi.* 2000; Griffiths *és mtsi.* 2006). A szabadban élő amőbák, a talaj- és a vízi ökoszisztémák fontos komponensei, egyre inkább elismert vektorai a változatos bakteriális eredetű humán patogéneknek (Everett 2000; Corsaro *és* Venditti 2004; Horn *és mtsi*. 2004; Greub *és* Raoult 2002; Griffiths *és mtsi*. 2006).

[e]Filogenetika: A filogenetika alapfeladata matematikai szemszögből jól definiált probléma: egy helyes fatopológiát kell hozzárendelnünk a különböző fajokból izolált, hasonló funkciójú és hasonló szekvenciájú (homológ) nukleinsavakhoz vagy fehérjékhez.



2. ábra. A *Chlamydiák* régi és új taxonómiájának összehasonlítása

1999-ben egy új, filogenetikai kapcsolatokon alapuló taxonómiai rendszer bevezetését javasolták, az új rendszerben több új nemzetséget és fajt vezettek be. (Bush *és* Everett 2001 nyomán, ábra átrajzolva).

A szakirodalomban nemrég jelent meg egy új módszer leírása, amely számos konzervált inszerciókból és deléciókból (azaz indelekből) álló molekuláris aláírást (signature), ritka genomi változásokat (*R*are *G*enomic *C*hange*s*, RGCs) használ, vagyis olyan különféle típusú fehérjéket, amelyek megkülönbözető jellegűek az összes elérhető *Chlamydia* fajra nézve és nem találták meg egyetlen másik baktériumban sem. A módszer egy más típusú taxonómiai marker leírása, mely olyan teljes, egész fehérjékből áll, melyek specifikusak a *Chlamydia* fajok csoportjaira, ezen csoportok azonosítására, valamint evolúciójuk és fiziológiai/élettani karaktereik megértésére alkalmas hatásos eszközt biztosítva ily módon (Gupta 2008; Griffiths *és mtsi*. 2006).

2.2. Az *Escherichia coli* törzsek

Az *Escherichia coli* kétségkívül egyike a leginkább tanulmányozott baktériumoknak, genetikai, patológiai és ipari szempontból is jelentős. Az ismert genomú *E. coli* törzsek számának növekedése egyedülálló lehetőséget kínál arra, hogy a mikroszatellita evolúciót ezekben a baktériumokban is tanulmányozzuk. A mikroszatellita-eloszlás vizsgálatban a hét *Chlamydia* teljes genom mellett négy *Escherichia coli* törzs (Blattner *és mtsi.* 1997; Hayashi *és mtsi.* 2001; Perna *és mtsi.* 2001; Welch *és mtsi.* 2002) teljes genomszekvenciáját is vizsgáltuk.

2.2.1. Az *Escherichia coli*ról

Az *Escherichia coli* fontos komponense a bioszférának. Az állatok alsó béltraktusát kolonizálja (Blattner *és mtsi*. 1997). Széleskörűen elterjedt az ember és a melegvérű állatok bélrendszerében. A bélben a túlsúlyban levő fakultatív anaerob szervezet, az alapvető bélflóra része, amely az egészséges gazda normális életműködését tartja fenn (Feng 2002a). Az emberi normál bélflóra általában szabályos sorrendben fejlődik ki a születést követően egy stabil baktérium populációhoz vezetve, ami a normál felnőtt bélflórát alkotja. Gram pozitív (*bifidobaktériumok*ból és *laktobacillusok*ból álló) populáció dominál kezdetben a gasztrointesztinális traktusban a korai élet során, amíg az újszülött szopik. Ez a bakteriális populáció redukálódik és némikébb kiszorítódik a Gram negatív flóra (*Enterobacteriaceae*) által, amikor az újszülöttet elválasztják (Baron 1996).

Az *E. coli* az *Enterobacteriaceae* család tagja, amely család sok nemzetséget- köztük olyan ismert kórokozókat mint a *Salmonella*, a *Shigella* és a *Yersinia* foglal magában (Feng *és mtsi.* 2002a). Fakultatív anaerob szervezetként a természetbe kikerülve is képes életben maradni és sokféle új gazdaszervezetbe bejutni. A patogén *E. coli* törzsek a bél-, vizeletkiválasztó, idegrendszeri, valamint a tüdő fertőzéseiért felelősek (Blattner *és mtsi*. 1997).

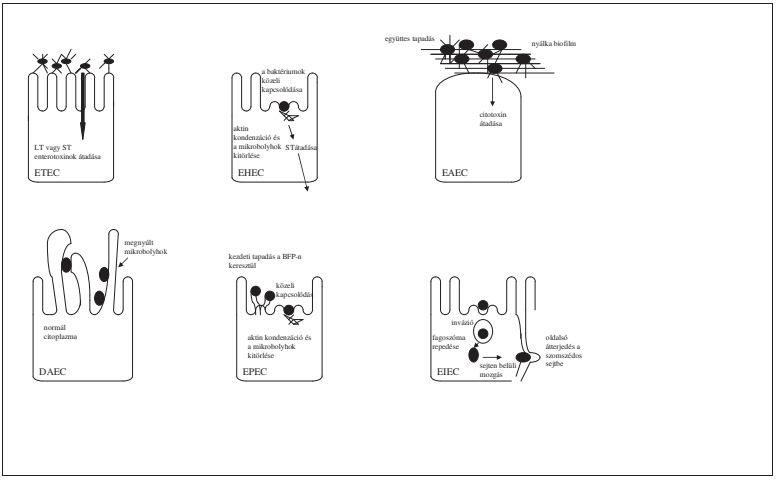
2.2.2. Az *Escherichia coli* törzsek által okozott betegségek

Az *Escherichia coli* (*E. coli*) az ember bélrendszerének uralkodó nem patogén fakultatív flóráját alkotja. Mindazonáltal néhány *E. coli* törzs olyan képességre tett szert, mellyel különféle humán megbetegedéseket okozhatnak.

Az *E. coli* O157:H7 több különböző/különféle betegséget okozhat, ezek közé tartozik a véres hasmenés és a hemolítikus urémiás szindróma (*H*emolytic *U*remic *S*yndrome, HUS). A betegség a vérlemezkék számának hirtelen csökkenésével, a vörösvértestek károsodásával és a veseműködés megszűnésével jár. Az O157:H7 szerotípust először 1983-ban írták le, miután 1982-ben egymás után két hemorrhágiás kolitisz járvány tört ki egy kínai gyorséttermi lánc oregoni és michigani éttermeiben. Az *E. coli* O157:H7 a Shiga toxin termelő *E. coli* csoportba tartozik (*S*higa *T*oxin *P*roducing *E. coli*, STEC). Az STEC jónéhány különböző *S*higa *t*oxint (Stx) termel és úgy tartják, hogy a HUS a Shiga toxinoknak a keringési rendszer endoteliális sejtjeire való hatásának, amely a Shiga toxinok szisztémás működéséből származik az eredménye. Az *E. coli* O157:H7 egy fontos táplálék által közvetített patogénként jelent meg. Az *E. coli* O157:H7 fő/elsődleges tárolóhelyei (rezervoárjai) az élelmezési célból tartott haszonállatok különösen a szarvasmarha, amely bélrendszerének távoli részében ad menedéket ezeknek a baktériumoknak. Az *E. coli* O157:H7 fekáliával fertőzött vízzel vagy étellel kerülhet át emberekbe (Lindstedt 2005).

A hasmenést, gyomorrontást, humán gasztroenteritiszt okozó *E. coli* törzseket hat csoportba sorolják (3. ábra): enteroaggregatív (*E*ntero*a*ggregative *E. coli*, EAEC), enteroemorrhágiás (*E*ntero*h*emorrhagic *E. coli*, EHEC)[f], enteroinvazív (*E*ntero*i*nvasive *E. coli*, EIEC), enteropatogén (*E*ntero*p*athogenic *E. coli*, EPEC), enterotoxikus (*E*ntero*t*oxigenic *E. coli*, ETEC), és a diffúz módon adheráló (*D*iffuse *A*dherent *E. coli*, DAEC). A patogén *E. coli*kat felszíni antigénjeik (O: szomatikus, H: flagelláris, K: kapszuláris) alapján sorolják be az egyes szerológiai csoportokba. Mind a hat felsorolt kategóriát különböző patogenezis és különböző O:H szerotípus- készlet jellemez (Nataro *és* Kaper 1998, Feng *és mtsi.* 2002b).

[f]EHEC: *Escherichia coli* O157:H7 EDL933 és *Escherichia coli* O157:H7 Sakai



3. ábra. A hasmenést okozó *E. coli* patogén vázlata

A hasmenést okozó *E.coli* mind a 6 azonosított kategóriája egyedi jellegzetességgel rendelkezik az eukarióta sejtekkel való interakciót illetően. Az ábra sematikusan mutatja be az egyes kategóriákba tartozó baktériumok és egy tipikus célsejt interakcióját. Meg kell jegyezni, hogy ezek a leírások főleg *in vitro* vizsgálatok eredményeiből származnak, meglehet, hogy nem teljesen tükrözik a fertőzött betegekben megjelenő jelenségeket (Nataro *és* Kaper 1998 nyomán, ábra átrajzolva).

Enteroaggregatív *E. coli* (EAEC)

EAEC fertőzés esetén jellemző a folyamatos, állandó vizes hasmenés (>14nap),melyet láz és/vagy hányás nem kísér, különösen a fejlődő országokban fordul elő.

Enterohemorrhágiás *E.coli* (EHEC)

Az EHEC erős toxint bocsát ki, melyeket verotoxinoknak vagy Shiga toxinoknak neveznek (az elnevezés onnan ered, hogy ezek a toxinok közeli hasonlóságot mutatnak a *Shigella dysenteriae* által kibocsátott Shiga toxinokkal). Ezeket a szervezeteket gyakran emlegetik STEC-ként (STEC, Shigatoxin-producing *E. coli*), azaz Shiga toxin termelő *E. coli*ként. Már nagyon alacsony infekciós dózis (100.-.200 organizmus) elégséges a betegség kialakulásához (Nataro *és* Kaper 1998).

Enteroinvazív *E. coli* (EIEC)

Az EIEC a bélrendszer hámsejtjeit támadja meg, vizes hasmenést okoz. A betegek kis hányadában az EIEC a shiggellózishoz hasonló tüneteket okoz. Az EIEC által okozott hasmenéses tünetek a fertőzött étel elfogyasztása után 12-72 órával jelentkeznek. A beteg állapota a következőkkel jellemezhető: nyálkás, véres széklet, hasi görcsök, hányás, láz, hidegrázás és általános rossz közérzet. A legtöbb dokumentált járvány étel okozta[g] vagy ivóvíz okozta betegség[h], bár a személyről személyre való átadódás is lehetséges. A járványok szennyezett hamburgerhús és pasztörizálatlan tej fogyasztásával voltak kapcsolatba hozhatók (Nataro *és* Kaper 1998; U.S. CFSAN 1992).

[g]Az étel okozta betegségek az étel elfogyasztása következtében alakulnak ki. Bár az étel okozta betegségeket általában tévesen ételmérgezésnek nevezik. Az igazi ételmérgezés akkor alakul ki, amikor valaki kémiai anyaggal vagy természetes toxinnal szennyezett ételt fogyasztott, míg az étel okozta betegségek többsége valójában olyan étellel kapcsolatos fertőzés, melyet egy sor ételben előforduló patogén baktérium, vírus, prion vagy parazita okozhat (U.S. CDC 2005). Ez a fajta fertőzés általában az élelemiszerek nem megfelelő kezeléséből, készítéséből és tárolásából származik. Az ételgyártás alatti, azt megelőző és követő megfelelő tisztálkodás (hygiené) csökkentheti a betegség „elkapásának" esélyét. Az élelmiszerek ellenőrzésén keresztül megvalósuló tevékenység mellyel az étel okozta betegségek kialakulását akadályozzák meg az élelmiszer biztonság. Étel okozta betegségek forrása lehet sokféle, változatos, a környezetet befolyásoló toxin.

Az étel okozta betegségek felelősek a populáción belül észlelhető magas morbiditás[g1] és mortalitás[g2] szintért, különösképpen a veszélyeztetett csoportokban (csecsemők, kisgyermekek, fiatalkorúak, idősek és az immunhiányban[g3] szenvedők). A WHO[g4] Élelbiztonsági, Zoonózis és Étel Okozta Betegségek Hivatala (Department of Food Safety, Zoonoses and Foodborne Diseases) annak érdekében, hogy az étel okozta betegségek előfordulását és gazdasági következményeit csökkenteni lehessen a tagállamokkal együttműködik programjuk létrehozásában és stabilizálásában/megerősítésében, annak érdekében, hogy biztosítani tudják az ételek biztonságát a gyártástól a végső fogyasztásig. Ebben a tekintetben a WHO páratlan kapacitást kínál fel, az egészség iránti elkötelezettségén keresztül, a kormánnyal, az élelmiszeriparral és a fogyasztókkal együttműködve azért, hogy jobban fókuszáljanak a nemzeti élelbiztonságot adó erőfeszítésekre és, hogy megerősítsék azokat (WHO 2007). A hasmenéses megbetegedések a fő okai a megelőzendő elhalálozásnak, különösen a fejlődő országokban élő öt év alatti gyermekek körében. Az elsőbbséggel bíró szükséges beavatkozásokat összegző tanulmányt (Dean *és mtsi*. 2006) a Világbank és az OUP (*O*xford *U*niversity *P*ress) együttesen jelentette meg.

[g1]morbiditás: 100 000 lakosból 1 év alatt hányan betegszenek meg

[g2]mortalitás: 100 000 lakosból 1 év alatt hányan halnak meg a kialakult betegség következtében

[g3]immunhiány: az immunrendszer csökkent (néha hiányzó működését) nevezzük immundefektusnak. Ha a defektus súlyos, a szövődmények is azok, az enyhébbek viszont tünetmentesek lehetnek. Erre az állapotra jellemző a fertőzésekre való fokozott fogékonyság.

[g4] WHO: Egészségügyi Világszervezet (*W*orld *H*ealth *O*rganization)

[h]Az ivóvíz okozta betegségeket olyan patogén mikroorganizmusok okozzák, melyek közvetlenül a szennyezett ivóvíz fogyasztásakor adódnak át. Az ételgyártás és ételkészítés során felhasznált szennyezett ivóvíz a forrása az étel okozta betegségnek, ugyanazon mikroorganizmus elfogyasztásán keresztül. A WHO szerint a hasmenéses megbetegedéseket a betegségben vagy rokkantságban eltöltött éveknek (*D*isability *a*djusted *l*ife *y*ears, DALY[h1]) megfelelő egyesített betegségi teher (*g*lobal *b*urden of *d*isease, GBD) 4,1%-ának becsülik és ezek felelősek 1,8 millió ember haláláért minden évben. Korábbi becslés szerint a teher 88%-a a nem biztonságos vízellátásnak, fertőtlenítésnek és hygienének tudható be és főként a fejlődő országok gyermekeit érinti. Az ivóvíz okozta betegségeket protozoák, vírusok, baktériumok és bélparaziták okozhatják (U.S. CDC 2005).

[h1]DALY: Betegségben vagy rokkantságban eltöltött évek (*D*isability *a*djusted *l*ife *y*ears,) a teljes egyesített betegségi terhet méri. Eredetileg a WHO fejlesztette ki

Enteropatogén *E. coli* (EPEC)

Az EPEC a legrégebben felismert hasmenéses megbetegedést okozó *E. coli* kategória. Az EPEC vagy vizes vagy véres hasmenést okoz (U.S. CFSAN 1992). A fejlődő országokban a csecsemőkori hasmenés fő okozója. Az anyatejes táplálás védettséget biztosíthat a betegséggel szemben, *in vitro* kísérletekkel igazolták, hogy az anyatej (és előtej, colostrum) erősen gátolja az EPEC Hep-2 sejtekhez való tapadását, adhézióját (Nataro *és* Kaper 1998).

Enterotoxigenikus *E. coli* (ETEC)

Az ETEC törzsek kétféle toxinnal „dolgoznak": vagy a hőlabilis toxinnal (LT), vagy a hőstabil toxinnal (ST), vagy mindkettővel. Az ETEC két fő klinikai szindrómával kapcsolatos, ezek: az utazók hasmenése és a harmadik világ gyerekeiben megjelenő „elválasztott hasmenés" (Nataro *és* Kaper 1998). A mikroorganizmusok a bélhártyát kolonizálják, itt fejezik ki enterotoxinjaikat. Az inkubációs periódus rövid (14-50 óra). A hasmenés vizes, általában vér és nyálka nélküli. Láz nem jelentkezik, a betegség általában önkorlátozó (U.S. CFSAN 1992) Az ETEC fertőzés a *Vibrio cholerae* által okozott kolerához hasonló hasmenést okozhat. Epidemiológiai vizsgálatokkal kimutatták, hogy a fécesszel szennyezett étel és ivóvíz a leggyakoribb hordozó közegei az ETEC fertőzésnek.

Diffúz módon adheráló E. coli (DAEC)

Keveset tudunk a DAEC-okozta megbetegedések epidemiológiájáról és klinikai profiljáról. Több tanulmány is a DAEC-t hozta kapcsolatba a hasmenéssel, míg más kísérletekből nem nyerték gyakrabban vissza a DAEC törzseket a hasmenéses betegekből, mint a tünetmentes kontroll személyekből (Nataro *és* Kaper 1998).

2.2.3. A coli csoport filogenetikája

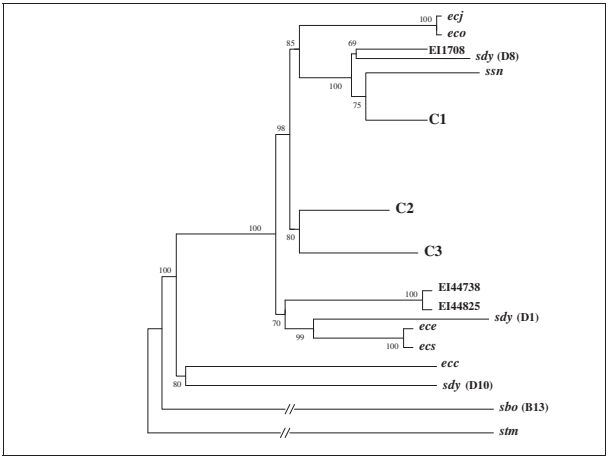
A legtöbb állat, beleértve az embert is, bélrendszere az *Escherichia coli*nak természetes élőhelye. Néhány *E. coli* törzs egy sor változatos bélrendszeri- (intesztinális) és bélrendszeren kívüli (extraintesztinális) betegséget (pl. hasmenés, a vizeletkiválasztó szervrendszer különféle fertőzései, vérmérgezés, csecsemő agyhártyagyulladás) okoz. A filogenetikai analízisek megmutatták, hogy az *E. coli* törzsek négy fő filogenetikai csoportba

(A, B1, B2 és D) tartoznak (Herzer *és mtsi.* 1990; Selander *és mtsi*. 1987; Clermont *és mtsi*. 2000), valamint azt, hogy a virulens bélrendszeren kívüli törzsek többsége főleg a B2 csoportba, a többi törzs pedig a D-be tartozik, míg a sokkal inkább kommenzalista törzsek az A csoportba tartoznak. Tulajdonképpen a filogenetikai csoportosítást vagy több lókuszos elektroforézissel vagy ribotipizálással lehet elkészíteni, mindkettő technika nagyon összetett és időigényes, ezen felül rendelkeznünk kell a tipizált törzsek gyűjteményével (Bingen *és mtsi*. 1998; Clermont *és mtsi*. 2000). Egyes becslések (Reid *és mtsi*. 2000) szerint az *Escherichia coli* K-12 és O157 (EHEC) törzsek kb. 4,5 millió éve divergálódtak közös ősüktől.

Az *E. colik* filogenetikájával kapcsolatban fontos megemlékezni a *Shigellák*[i]kal való rokonságról. Már régóta tudjuk, hogy a *Shigellák* közeli rokonságban állnak az *Escherichia colik*kal. Már korábban leírták, hogy az EIEC törzsek ugyanazokat a virulenciával kapcsolatos géneket hordozzák, mint a *Shigellá*k (Hsia *és mtsi.* 1993), és az enteroinvazív jelleget több *E. coli* szerológiai csoportban, mindazonáltal ezen csoportoknak nem minden tagjában fejeződik ki az invazív fenotípus (Pál *és mtsi*. 1997).

Mindent összevetve kevés olyan megkülönböztető jegy van, ami alapján a *Shigella* törzsek elkülöníthetők a szintén vérhast okozó EIEC törzsektől. Sőt a mai filogenetikai tanulmányokban azt javasolták, hogy a *Shigella* és az EIEC alkossák az *E. coli* egyik patovarját (Lan *és mtsi.* 2004; Yang *és mtsi.* 2007). Yang *és mtsi*. 23 kromoszómális gén alapján szerkesztettek egy filogenetika fát, amely megmutatja, hogy a *Shigella* törzsek többsége három fő *Shigella* klaszterbe (C1, C2, C3) esik, A három EIEC törzs nem sorolható egyik *Shigella* klaszterbe sem, hanem független ágakon találhatók. A legtöbb elágazódás magas bootstrap értékei (lásd 27. oldal) megerősítik a fa megbízhatóságát. Ezen fa segítségével mutatom be, kontextusba helyezve, a vizsgált *E. coli* törzsek filogenetikai kapcsolatát (4. ábra).

[i]A *Shigella* fajok elsődleges okozói (etiológiai ágensei) a bakteriális vérhasnak vagy más néven shigellózisnak, ami komoly fenyegetést jelent a közegészségre nézve, különösen a kevésbé fejlett országokban, ahol a fertőtlenítés nagyon kezdetleges és szegényes. 160 millióra becsülik az évenként előforduló shigellózisok számát világszerte, ezek a megbetegedések közül 1,1 milló végződik halállal, főként az 5 éven aluli gyermekek körében (Kotloff *és mtsi*. 1999, Yang *és mtsi.* 2007) A *Shigellák*at először egy japán kutató, Shiga fedezte fel az 1890-es években és genusként 1950-ben fogadták el. A biotipizálás alapján a genust 4 fő fajra osztották: *S. dysenteriae*, *S. flexneri*, *S. boydii*, *S. sonnei* (Hale 1991, Yang *és mtsi.* 2007).



4. ábra. A *Shigella* klaszterek és az *Escherichia coli* törzsek filogenetikai kapcsolata

A filogenetikai fa szomszéd összevonó (NJ, Neighbor Joining) módszerrel készült 23 kromoszómális gén egyesített adataiból. Az EI rövidítésű törzsek az EIEC törzsek, a rövidítés után a törzs száma áll. A két nempatogén *E. coli* törzs: *ecj*: *Escherichia coli* K-12 W3110 (GenBank: AP009048) és *eco*: *Escherichia coli* K-12 MG1655 (GenBank: U00096). A két enterohaemorrhágiás coli (EHEC) törzs: *ece*: *Escherichia coli* O157:H7 EDL933 (GenBank: AE005174) és *ecs*: *Escherichia coli* O157:H7 Sakai (GenBank: BA000007). Az uropatogén *E. coli* (UPEC) törzs: *ecc*: *Escherichia coli* CFT073 (GenBank: AE014075). Az ábrán szerepő *Shigella* törzsek: *sdy*: *Shigella dysenteriae* (GenBank: CP000034), *ssn*: *Shigella sonnei* (GenBank: CP000038), *sbo*: *Shigella boydii* (GenBank: CP000036). A rövidítéseket a szerotípusszám követi, pl.: *sdy* (D1). A *Shigella* törzsek három fő klaszterét C1, C2, C3 jelöli. Az 50% feletti bootstrap értékeket (lásd 27. oldal) a elágazódásoknál (nodusoknál) tüntették fel. A *Salmonella typhimurium*ot választották külcsoportnak, *stm*: *Salmonella typhimurium* LT2 (GenBank: AE006468), Yang *és mtsi*. 2007 nyomán, ábra átrajzolva).

3. Célkitűzések

Horizontális géntranszferrel (HGT) a genomok távoli rokon szervezetekből szerezhetnek géneket, ezáltal a HGT a baktériumok genetikai diverzitásának egyik fő forrása lehet. Kérdés, hogy milyen mértékben vesz részt a HGT a genomszerkezet változtatásában, meg lehet-e határozni a HGT előfordulását és irányát és vajon el lehet-e dönteni egy teljes genom összes génjéről, hogy HGT eseménnyel érkezett-e a genomba. Célunk volt olyan általánosan elfogadható módszert találni, mely alkalmas a HGT megbízható kimutatására. Genomi jellemzőik (pl. viszonylag kis genom méret) és orvosbiológiai jelentőségük alapján választottuk a hozzáférhető 5[a] *Chlamydia* törzs teljes genom szekvenciáját egy HGT eseményeket szisztematikusan kimutató, bioinformatikai módszereket alkalmazó, összehasonlító analízishez.

A genomevolúció egy másik igen érdekes aspektusa az egyszerű szekvenciális ismétlődések (*S*imple *S*equence *R*epeats, SSR-ek), másnéven mikroszatelliták vizsgálata. Az SSR-eket főleg eukariótákban vizsgálták. Napjainkra (2008. február) több mint 600 teljes mikrobiális genomszekvencia található az Entrez Genome adatbázisban (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1). A hozzáférhető szekvenciák mellett a mikroszatellita evolúció becslésére vonatkozó új technikák megjelenése is lehetővé teszi az SSR-ek részletes vizsgálatát prokarióta genomokban is. További célul tűztük ki a *Chlamydiales* rend 7 tagjában, valamint 4 *Escherichia coli* törzsben a mikroszatelliták eloszlásának vizsgálatát. A két választott csoport közül a *Chlamydiák* monofiletikus csoportja filogenetikai szempontból jól elkülönül a többi taxontól (Stephens *és mtsi*. 1998; Kalman *és mtsi*. 1999; Read *és mtsi*. 2000, 2003; Shirai *és mtsi*. 2000; Chen *és mtsi*. 2007). Genomevolúciójukat nemrég tanulmányozták bioinformatikai módszerek segítségével (Ortutay *és mtsi*. 2003; McNally *és mtsi*. 2007). Ezen tulajdonságaik alapján ideális célpontjai lehetnek egy összehasonlító analízisnek. Az *Escherichia coli* genomokat választottuk másik célcsoportnak, mert ezek a baktériumok a legjobban tanulmányozott prokarióták közé tartoznak, jól leírt genetikával.

[a]vizsgálataink megkezdésekor (2001-ben) érvényes adat

Az ismert genomú *E. coli* törzsek növekvő száma kivételes lehetőséget kínál a mikroszatellita evolúció analízisére ezekben a közel rokon baktériumokban (Blattner *és mtsi*. 1997; Hayashi *és mtsi*. 2001; Perna *és mtsi*. 2001; Welch *és mtsi*. 2002). A 4 *Escherichia* genomot egyfajta referenciaként használtuk más kutatásokhoz hasonlóan (Azad *és mtsi*. 2007; McNally *és mtsi*. 2007).

Az általános trendek megállapításán felül célunk volt továbbá az is, hogy összehasonlító bioinformatikai analízis alkalmazásával újabb információkhoz jussunk mikroszatellitákban gazdag génekről, hogy jobban megértsük a mikroszatelliták prokarióta genomokban betöltött szerepét.

4. Az alkalmazott módszerek

4.1. Horizontális géntranszfer kimutatására alkalmas módszerek

HGT események detektálására háromféle módszert alkalmaznak áétalánosan. Az első az organizmus összes génjével végzett szisztematikus hasonlósági keresés egy kellően nagy adatbázissal szemben. A találati listában a keresett génnel legnagyobb hasonlóságot mutató találatok közül olyanokat kell keresnünk, amelyek filogenetikailag nem rokon taxon(ok)ból származnak (Faguy *és* Doolittle 1999).

A második módszer a filogenetikai analízis. Ha egy gén HGT eseménnyel került a genomba (HGT gén), a HGT génre és rokonaira készített filogenetikai fa topológiáját vizsgálva azt találjuk, hogy a HGT gént befogadó genom nem a filogenetikai rokonságához csatlakozik, hanem annak a genomnak a rokonságához, ahonnan a HGT gén származik (Stanhope *és mtsi.* 2001).

A harmadik módszer a kodon adaptációs index (*C*odon *A*daptation *I*ndex, CAI) (Sharp *és* Li 1987) és a génszekvenciák százalékos GC tartalmának vizsgálatával történő HGT kimutatás. A kodon adaptációs index (CAI) egy 0 és 1 közé eső mérőszám ($0 \leq CAI \leq 1$), amely a kodonhasználatra vonatkozó relatív alkalmazkodást fejezi ki, oly módon hogy az adott génre jellemző kodonhasználatot egy nagy mértékben expresszálódó génkészlet kodonhasználatához viszonyítja. Minél nagyobb a CAI, annál jobban „simul" az adott gén a genomba. A közelmúltban lezajlott HGT eseményből származó génekről az feltételezzük, hogy még nem volt idejük alkalmazkodni, illeszkedni (az „ameliorate" angol nyelvű kifejezésből) a genomhoz, ezért feltehetően viszonylag kis CAI értékekkel rendelkeznek. A százalékos GC tartalom-eltérést kimutató módszer azon a feltételezésen alapul, hogy adott genomban a géneknek van egy jellemző átlagos GC tartalma és létezik egy a genomra jellemző kodonhasználati mintázat. Ha a gén a közelmúltban HGT-vel érkezett a genomba, akkor ezek a jellemző értékek nem a gazda genom értékeire fognak hasonlítani, hanem annak a genomnak a megfelelő jellemző értékeire, ahonnan az adott gén származik (Lawrence *és* Ochman 1997, 1998).

Vizsgálataink során mindhárom módszert alkalmaztuk, a hasonlósági keresés és a filogenetikai analízis mellett vizsgáltuk a kodon adaptációs index (CAI) és a GC tartalom HGT-re vonatkozó diagnosztikus jelentőségét is.

4.1.1. Hasonlósági keresés a FASTA3 és BLAST programcsomaggal

A vizsgálatba bevont *Chlamydia* genomok valamennyi fehérjét kódoló nyílt leolvasási keret (*O*pen *R*eading *F*rame, ORF) szekvenciával folytattunk hasonlósági keresést FASTA3 (Pearson 1999) és BLAST (Altschul *és mtsi*. 1997) programcsomag segítségével. A FASTA programmal végzett hasonlósági keresésnél egy lokális adatbázissal - mely tartalmazta az akkor elérhető[a] valamennyi bakteriális és archaeális ORF fehérje átiratát -szemben kerestünk. Az E érték, (*E*xpected Value) küszöbe 10^{-10} volt, tehát az ennél nagyobb E értékű találatokat a továbbiakban nem vettük figyelembe. (Az E érték azt a valószínűséget fejezi ki, hogy a kapott találat pontszáma a véletlen következménye). E mellett a teljes nemredundáns fehérje adatbázissal[b] szemben végeztünk standard BLAST hasonlósági keresést TBlastn program felhasználásával. Az E értéket ebben az esetben is 10^{-10}-nél határoztuk meg.

4.1.2. Filogenetikai analízis: Az alkalmazott filogenetikai módszerek elméleti áttekintése

Az elvégzett filogenetikai vizsgálatok ismertetése előtt szükségesnek tartom az alkalmazott módszerek rövid ismertetését a könnyebb érthetőség végett.

4.1.2.1. A maximális parszimónia (MP, Maximum Parsimony) módszer

A maximális parszimónia (MP, Maximum Parsimony) módszer a karakter alapú módszerek közé tartozik. A karakter alapú módszerek megtartják a taxon eredeti szekvenciáját, ezzel rekonstruálhatóvá téve a nóduszok, vagyis az ősi fajok lehetséges szekvenciáját. A törzsfa elkészítéséhez nemcsak a szekvenciák közötti távolságra, hanem a szekvenciákban levő információkra is szükség van. A MP célja az evolúciós fa ágai összhosszúságának minimalizálása. Olyan fát keres, amely a lehető legkevesebb evolúciós lépéssel (karakterváltozással, ill. feltételezett mutációval) éri el a leszármazási viszonyok megmagyarázását. Számos azonos pontszámú fát szolgáltat, ezek közös részét vehetjük, mint megbízhatót. Nagy távolságú szekvenciák esetében hátránya, hogy azonos bázis esetén azt tételezi fel, hogy nem történt mutáció, holott jelentős a telítődés, visszacserélődés vagyis a back mutáció valószínűsége.

[a]Letöltve 2001. február 19., 27 bacteria és 8 archaea teljes genom. A dolgozat írásának időpontjában (2008. április) 629 bakteriális és 53 archaeális eredetű teljes mikrobiális genomszekvencia található az Entrez Genome adatbázisban (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1).

[b] Letöltve 2002. január 4.

Az illesztett szekvenciák egy taxontulajdonság mátrixnak tekinthetők, az egyes pozíciók a jellegnek, az adott pozíción szeplő nukleinsav pedig a karakter állapotának felelnek meg. A tulajdonságok közös eredetén alapuló valódi hasonlóságot homológiának nevezik. Ha egy jellegállapot két faj között hasonló, de valójában csak a konvergens evolúció következménye, akkor azt homopláziának vagy analógiának hívjuk. A maximális parszimónia módszere szerint tehát arra kell törekednünk, hogy a fa a legtöbb homológiát és a legkevesebb homopláziát tartalmazza (Podani 2003).

4.1.2.2. A szomszéd összevonó (NJ, Neighbor Joining) módszer

A szomszéd összevonó (NJ, Neighbor Joining) módszer a távolság alapú molekuláris filogenetikai módszerek közé tartozik. Ezekre a módszerekre jellemző, hogy először minden OTU[c]-t minden OTU-val páronként összehasonlítva valamilyen módszer alapján a szekvenciákból vagy távolság-mátrixot számolnak, és ebből kiindulva jutnak el a törzsfához. Ha két szekvencia közt több a különbség, akkor ahhoz a távolságmátrixban nagyobb számot rendelünk, mint ahhoz a szekvencia párhoz, melyek között kicsi a különbség. Ha két szekvencia között nincs különbség, akkor 0 a távolságuk (átló). Hátránya, hogy a szekvencia adatok távolsággá transzformálása információvesztéssel jár. Előnye, hogy kisebb a számítás igénye, mint a karakter alapú módszereknek. A módszer a belső águosszúságok minimalizálására törekszik, gyakorlatilag a maximális parszimónia módszert alkalmazza távolság adatokra. A Neighbor joining módszer a fában egymás mellé kerülő OTU-k kiválasztásánál nemcsak a hozzájuk tartozó távolságmátrix értéket, hanem az OTU-knak az összes többivel alkotott távolságát is figyelembe veszi. A módosított távolság annál kisebb lesz az eredetinél, minél nagyobb a két OTU átlagos távolsága a többitől, mert a nagy átlag (gyors evolúció) megnöveli a két OTU relatív közelségét. A módszer előnye, hogy gyors és egyszerű, de nem garantálja, hogy megtaláljuk az optimális fát. Eredményként egy gyökér nélküli fát kapunk (Podani 2003).

[c] *O*perational *T*axonomic *U*nit: kezelendő taxonómiai egység. Leggyakrabban fajok vagy magasabb taxonok, de lehetnek egyedek vagy populációk is, attól függően, hogy mi az osztályzás célja.

4.1.2.3. A maximum-likelihood („legnagyobb valószínűség") módszere

A maximum likelihood eljárás alkalmazása már egy konkrét evolúciós modell alkalmazását igényli, az evolúciós mintázat feltárásához pontosan meg kell adnunk, hogy miképpen alakulhatott át az egyik szekvencia a másikba. A maximum-likelihood módszer a modell ismeretében megadja, hogy a sok lehetőség közül melyik fa kialakulása a leginkább valószínű (Podani 2003). Az elemzést többféle helyettesítési modell mellett is elvégezhetjük (lásd nukleotid szubsztitúciós modellek). A számítások során a teljes szekvenciát figyelembe kell vennünk, nemcsak az eltéréseket okozó pozíciókat. A gyakoriságokból és *k* tranzíció/transzverzió hányadosokból meghatározható a modell „szíve" egy 4×4-es mátrix, amely a *nukleotidcserék* rátáit tartalmazza az evolúciós időegységre vonatkoztatva. A mutációs ráták segítségével kiszámítható annak az eseménynek a valószínűsége, hogy *t* idő elteltével mondjuk az *A* bázis helyére a *G* bázis kerül. Ezt a valószínűséget $P_{AG}(t)$-vel jelöljük. Annak az esélye (*L*= likelihood) például, hogy adott szekvencia valamely pozíciójában az *A* nukleotid van, s ezt t idő elteltével *G* váltja fel, a következő:

$$L_{AG}(t) = f_A P_{AG}(t)$$

ahol f_A az *A* nukleotid gyakorisága a kezdeti szekvenciában. Ha feltételezzük, hogy a szekvencia minden egyes pozíciója függetlenül változik a többitől az evolúció során, akkor annak az esélye, hogy X szekvenciából *t* idő elteltével éppen az Y szekvenciát kapjuk a következő likelihood-függvénnyel kapható meg:

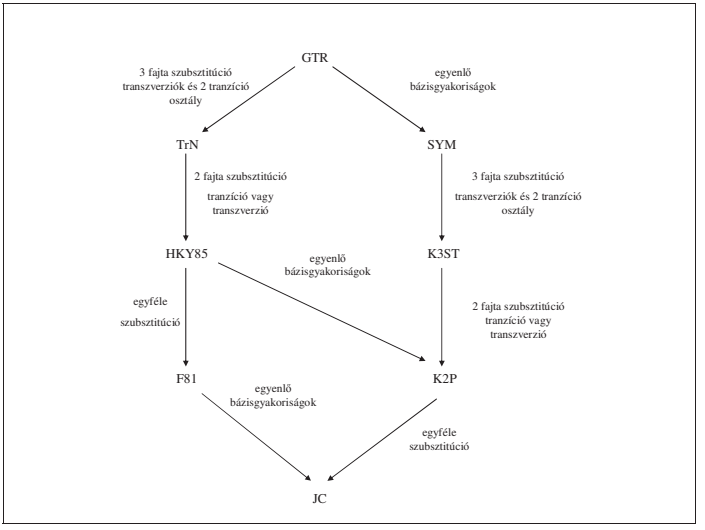
$$L_{XY}(t) = \prod_{i=1}^{S} f_{x_i} P_{x_i,y_i}(t)$$

ahol *s* a két szekvencia hosszúsága, x_i és y_i pedig az *i*-edik pozícióban található nukleotid az X, ill. az Y szekvenciában. Miután ez rendszerint igen kicsi szám, célszerű a $\ln L_{XY}(t)$ átalakítás, így a számítások is jelentékenyen leegyszerűsödnek. A feladat egy olyan fa megtalálása, amelyre a szorzat maximális. Ez a fa mutatja a legvalószínűbb leszármazási mintázatot, feltéve, hogy a modell kiindulási feltételei helyesek voltak (Podani 2003).

4.1.2.4. Nukleotid szubsztitúciós modellek

A maximum likelihood módszerrel működő program minden pozícióra kiszámítja, hogy adott fa és helyettesítési modell mellett mi a valószínűsége annak, hogy a megfigyelt variációs mintázat jöjjön létre az adott pozícióban. Az egyes pozíciókra kapott valószínűségek összeszorzásával adódik a teljes fa valószínűsége. Ezt a program sok fára megnézi és a legjobbat kiválasztja. Az elemzést többféle helyettesítési modell mellett is elvégezhetjük, ezek közül is kiválasztva a legjobbat.

Az általános modellre (*G*eneral *T*ime *R*eversible, GTR, Rodrígez *és mtsi.* 1990) egyre több megkötést téve, fokozatosan eljuthatunk a Jukes-Cantor modellhez (Jukes *és* Cantor 1969), (5. ábra). Ha egyenlő egyensúlyi gyakoriságokat tételezünk fel, a SYM modellt kapjuk (Zarkikh 1994). Erre továbbfeltételezve, hogy a tranzíciók rátája, valamint az A↔T és C↔G, illetve az A↔C és G↔T transzverziók rátája megegyezik, Kimura három paraméteres modelljét kapjuk (Kimura 1981). Ha minden transzverzió rátáját azonosnak tekintjük, Kimura két paraméteres modelljéhez jutunk (Kimura 1980), végül a tranziciók rátáját a transzverziók rátájával egyenlővé téve eljutunk a Jukes-Cantor modellhez. Az általános modellből a Tamura-Nei modellhez jutunk, ha feltesszük, hogy minden transzverzió rátája azonos, *a=c=d=f*, (Tamura *és* Nei 1993). A Hasegawa-Kishino-Yano modellben az összes tranzíció rátája is azonos (Hasegawa *és mtsi.* 1985). Felsenstein 1981-es modelljében (Felsenstein 1981) minden szubsztitúciós ráta azonos típusú, de az egyensúlyi gyakoriságok különbözőek (Miklós 2002).



5. ábra. A reverzibilis nukleotid szubsztitúciós modellek családja

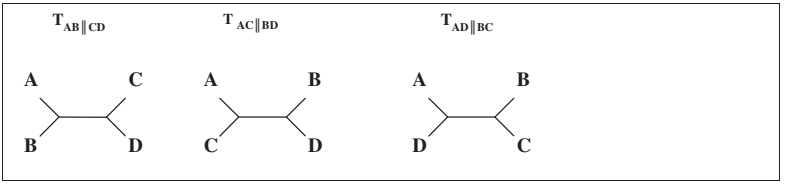
Az általános modellből (GTR, Rodrígez *és mtsi.* 1990) egyre több megszorítással jutunk el a Jukes-Cantor modellig (JC, Jukes *és* Cantor 1969). További rövidítések: TrN: (Tamura *és* Nei 1993), HKY85: (Hasegawa *és mtsi*. 1985), F81:(Felsenstein, 1981), SYM: (Zharkikh 1994), K3ST: Kimura 3 szubsztitúció-típusú modellje (Kimura 1981), K2P: Kimura 2 paraméteres modellje (Kimura 1980). (Hillis *és mtsi*. 1996 *és* Miklós 2002 nyomán, ábra átrajzolva).

4.1.2.5. A kvartett kirakó (*quartet puzzling*) algoritmus

Adott n elemű illesztett nukleotid (vagy aminosav) szekvenciakészlet bármely 4 tagú csoportját kvartettnek nevezzük. A kvartett kirakó algoritmus megvizsgálja/elemzi az adatbázis összes lehetséges kvartettjét, kihasználva azt, hogy a kvartettek számára csak 3 gyökér nélküli fa[d] topológia lehetséges (6. ábra).

[d]Az objektumok számát (m) tartalmazó általános formula $\prod_{i=3}^{m}(2i-5)=\frac{(2m-5)!}{2^{m-3}(m-3)}$

felhasználásával, amely a gyökérrel nem rendelkező fák lehetséges számát adja meg (Podani 2003), kiszámíthatjuk, hogy m=4 esetén 3 gyökér nélküli fa lehetséges.



6. ábra. A kvartettre vonatkozó 3 különböző informatív fa topológia

A q= (A,B,C,D) kvartett (m=4) esetén 3 gyökér nélküli fa lehetséges (magyarázat a szövegben). T (tree) az egyes fákat jelöli (Salemi 2003 nyomán, ábra átrajzolva).

Az algoritmus egy három lépésből álló eljárás. Az első lépésben, az ún. maximum-likelihood lépésben, generálja az összes lehetséges kvartettet: $\binom{n}{4} = \frac{n!}{4!(n-4)!}$.

A második lépésben, a kvartett kirakó (*quartet puzzling*) lépésben, a program közbenső fát generál, úgy, hogy egy már rekonstruált al-fába (sub tree) szekvenciákat illeszt. Ezt a lépést legalább 1000-szer ismétli a program, a szekvenciák különböző bemeneti sorrendjeivel, azért hogy elkerülje a szekvenciák sorrendiségéből származó/adódó melléktermékek rekonstruálását és annak érdekében, hogy a generált fáknak reprezentatív gyűjteményét kapjuk. Végül, a harmadik lépésben, a közbenső eredményfákból többségi konszezus elve alapján egy konszenzusfa készül, amit kvartett kirakó, azaz *quartet puzzling* fának nevezünk. A konszenzus lépés adott csoportnak a közbenső fákban való megjelenéséről ad információt. Ez az ún. megbízhatósági érték (reliability value), melyet %-ban adnak meg/fejeznek ki, azt méri, hogy milyen gyakran jelenik meg egy adott szekvencia csoport a közbenső fákban. Minden olyan csoport, amely a készlet több mint 50%-ában megjelenik, ábrázolva lesz a többségi konszenzus fában, amely fa nem feltétlenül maximum likelihood fa. A megbízhatósági érték nem keverendő a szokásos bootstrap értékkel (lásd 27. oldal), minthogy a megbízhatósági érték belső eredménye a kvartett kirakó algoritmusnak, míg a bootstrap elemzés egy külső eljárás, amit bármely faépítő módszerre lehet alkalmazni (Salemi 2003), viszont interpretálható ugyanúgy: az adott elágazás megbízhatóság mérőszámaként.

4.1.2.6. Az eredményfák kiértékelése

Az eredményfák kiértékelése, azaz a kapott fa alátámasztottságának tesztelése ún. újramintavételezési statisztikai módszerekkel (pl. bootstrapping) történhet („pull yourself up by your own bootstrap", azaz „Segíts magadon"). Lényege a mintából újra mintát veszünk, vagyis az eredeti szekvencia-illesztésből kiindulva az egyes pozíciókat visszatevéses mintavételezéssel megkeverjük és az így kapott mesterséges szekvenciák alapján újra kiszámítjuk a törzsfát. Ha a bootstrappelést sokszor ismételjük, akkor sok egymástól valamilyen szinten eltérő törzsfát kapunk. Ha a kapott törzsfákból konszenzus fát készítünk, és az ágakra ráírjuk, hogy hány százalékban támogatták a bootstrap fák az adott elágazást, akkor a kiindulási adataink, vagyis a kiválasztott szekvenciák jóságát becsülhetjük meg, azaz, hogy a szekvenciák mennyire voltak alkalmasak egy filogenetikai elemzés elvégzésére. Ha több 50%-nál kisebb támogatottságú ágat kapunk a konszenzus törzsfán, akkor tanácsos a vizsgált fajok más szekvenciáit is elemeznünk és az analízis kezdődhet elölről (Podani 2003).

4.1.3. Filogenetikai analízis: Az elvégzett vizsgálatok

A filogenetikai analízis első lépéseként a `FASTA3` hasonlósági keresés eredményei közül kiválasztottuk az első öt nem *Chlamydia* homológ találatot a további analízis számára. A fehérje szekvenciákat `ClustalW` programmal (Thompson *és mtsi*. 1994) illesztettük. Az illesztett fehérjeszekvenciák alapján elkészítettük a megfelelő kódoló nukleotid szekvenciák illesztését is. A filogenetikai fákat a nukleotid szekvencia illesztések alapján `TREE-PUZZLE` programmal *quartet puzzling* algoritmussal (Strimmer *és* von Haesler 1996; Schmidt *és mtsi.* 2000) készítettük. A *quartet puzzling* algoritmus a maximum likelihood faépítő módszer egyik változata, a program által felkínált több nukleotid szubsztitúciós modell közül a HKY nukleotid szubsztitúciós modellt (Hasegawa-Kishino-Yano 1985) választottuk.

Vizsgálatunk során az eredményfák topológiáját vizsgálva monofília töréseket kerestünk, azaz akkor fogadtuk el, hogy feltételezett HGT eseményt detektáltunk, amikor az adott filogenetikai fában a szorosan kapcsolt öt *Chlamydia* ágba idegen genomból kerül homológ. Ezután a `BLAST` hasonlósági keresések eredményei közül is kiválasztottuk azokat a homológ géncsoportokat, ahol az első öt találat között nem *Chlamydia* -ból származó találat is szerepelt. Ezeket a szekvenciákat is `ClustalW` programmal illesztettük.

Az illesztett szekvenciákból PAUP* programmal (Swofford 2002) *maximális parszimónia* (MP, maximum parsimony) algoritmus felhasználásával filogenetikai fákat szerkesztettünk, bootstrap analízist végeztünk.

A kiválasztott fák esetében további fákat szerkesztettünk PAUP* programmal, szomszéd összevonó módszerrel (NJ, Neighbor Joining) és maximális parszimónia módszerrel, bootstrap analízissel. Minden NJ és MP fát 1000-szer generáltunk (a NJ fáknál a p-távolságot használtuk). A bootstrap fákból konszenzus fák[e] születtek. A HGT-gyanús esetekben fehérje-szintű illesztést is készítettünk, ami nagyobb evolúciós távolságok (különböző doménekbe tartozó taxonok) esetén megfelelő. Végül az eredményként kapott adatokra a TREE-PUZZLE programmal is megkerestük a *quartet puzzling* fákat azokban az esetekben, ahol nem ez volt a kiindulási alap (BLAST hasonlósági keresések eredményei). Az aminosav szekvenciák esetén a távoli rokon aminosav szekvenciákra alkalmazható szubsztitúciós modellt, a WAG modellt (Whelan *és* Goldman 2001) választottuk. A három különböző filogenetikai módszer eredményfáiból konszenzusfák készültek, melyeket együtt vizsgálva választottuk ki azokat, ahol a monofíliatörés nem csak az adott módszerrel, hanem több módszerrel volt kimutatható. Itt kell megjegyezni, hogy a TREE-PUZZLE-t a NIIF (*N*emzeti *I*nformációs *I*nfrastruktúra *F*ejlesztési *I*ntézet) szuperszámítógépein futtattuk, ezzel abban az évben (2003) a legnagyobb számítógépes kapacitást kötöttük le, megelőzve minden más tudományterületet.

4.2. HGT események detektálása a kodonhasználat és a kodon-eloszlás vizsgálatával

A Lawrence *és* Ochman által alkalmazott módszer (Lawrence *és* Ochman 1998) módosított változatát alkalmaztuk. Az adott genom valamennyi ORF szekvenciájára kiszámoltuk a százalékos GC tartalmat és a kodonhasználat alapján a génre jellemző kodon adaptációs indexet (CAI). Kiszámoltuk továbbá az egész genomra jellemző százalékos GC arányt. Ezeket az értékeket vizsgáltuk valamennyi olyan gén esetében, amelynél szóba került a HGT érintettség a másik két módszer esetén. Akkor fogadtuk el adott gén esetében, hogy a módszer HGT eseményt valószínűsít, ha a gén százalékos GC tartalma legalább 5, de inkább 10 százalékkal eltért a genomi átlagtól.

[e](Nelson-féle) többségi-szabály konszenzus (majority rule consensus) fa: Megengedi bizonyos ellentmondó elágazások felbukkanását, amennyiben ezek a kiindulási törzsfák több mint 50%-ában szerepeltek. (a %-ot változtathatjuk). Előnye, hogy biztosan nem szerepel olyan elágazás a konszenzus fán, amely az egyik eredeti fában sem volt megtalálható (Podani 2003).

4.3. Mikroszatellita-eloszlás vizsgálata

4.3.1. A mikroszatelliták kiválasztása és osztályba sorolása

A vizsgálatba bevont genomokból a tökéletes ismétlődések kiválasztása egy saját PERL (Wall *és mtsi*. 2000) program segítségével történt. A maximum 6 bázispár (bp) hosszúságú ismétlődési egységeket tartalmazó minimum 12 bp hosszúságú szakaszokat vettük figyelembe (ld. Gáspári *és mtsi.* 2007). A nem tökéletes ismétlődéseket a *T*andem *R*epeats *F*inder programmal (TRF, Benson 1999) azonosítottuk.

Valószínűleg a TRF a tandem ismétlődések azonosítására/detektálására használt legnépszerűbb algoritmus. Például a Nemzetközi Humán Genom Szekvenáló Konzorcium (*I*nternational *H*uman *G*enome *S*equencing *C*onsortium, IHGSC) is a TRF-et használta a humán genomban található mikroszatelliták azonosítására (detektálására). A TRF végigpásztázza a szekvenciát hogy meghatározza azokat a régiókat, ahol a motívumok periodikusan ismételtek. A módszer egy statisztikai szabálykészlet alkalmazásán alapul, melyet az eredeti cikk ír le (Benson 1999). A legmegfelelőbb/legalkalmasabb motívumot azután meghatározza minden régióra és ezt a motívumot illeszti a régió mentén egy dinamikus programozású algotritmus (*W*raparound *D*ynamic *P*rogramming, WDP) használatával (Fischetti *és mtsi*. 1993). A WDP eljárás inputként (bemenetként) egy motívumot és egy szekvenciát vesz, egy optimális globális illesztést szolgáltat a szekvencia és a motívum tökéletes tandem ismétlődése között. A WDP optimalizálja mind az illesztési pontértéket (alignment score), mind pedig a motívum ismétlődéseinek a számát. A pontértéket a program számolja ki az illesztésből úgy, hogy pozitív súlyozást kap minden helyesen illesztett nukleotid (match) és negatív súlyozást kapnak a szubsztitúciók (mismatch) valamint az inszerciók és deléciók (indel) vagy rések (gap). Ha az illesztési pontérték magasabb egy küszöbértéknél (ezt a felhasználó állíthatja be). Különböző motívumokat lehet illeszteni ugyanazon régió mentén, mely esetekben csak a legjobb három detektálás tér vissza. Megjegyzendő, hogy a legjobb illesztés(ek) rövidebbek lehetnek mint a kezdetben detektált régió (Leclercq *és mtsi*. 2007).

A következő paraméterekkel dolgoztunk: egyezési érték (match score): 2, hibás pár érték (mismatch score): 3, indel érték (indel score):7; egyezési százalék (match percentage): 80, indel százalék (indel percentage): 10, az ismétlődő egység maximális hossza (maximum repeat length): 500, a legkisebb megengedett találati érték (minimum allowed score): 24. Ez a

beállítás azért volt szükséges, hogy azonosítani lehessen a 12 bp hosszúságú tökéletes ismétlődéseket, azaz, hogy ugyanazt a minimálisan detektálható/azonosítható hosszt kapjuk, mint a saját programunkat használva.

A nem tökéletes ismétlődésekben, ahol a `TRF` részleges ismétlődéseket is engedélyez (pl. 8,5 egység) az ismétlődések végén, szükségesnek tartottuk a részleges ismétlődések levágását (példánkban 8-ra kerekítve). Erre azért volt szükség, hogy jobban összevethetők legyenek a tökéletes ismétlődésekkel, ahol az ismétlődés mindig teljes egységekből áll (1. táblázat). Az SSR-eket az ismétlődő egységek és a szekvencián belüli elhelyezkedés alapján osztályoztuk (minden szekvencia, kódoló és nem kódoló régió szekvenciák (GenBank genomi szekvencia fájlok annotációs részében található CDS alapján). Az ismétlődő egység osztályok egységesítése (standardizálása) a korábbi leírás alapján (Jurka *és* Pethiyagoda 1995; Tóth *és mtsi*. 2000; Gáspári *és mtsi*. 2007) történt, például az „**`acg`**" osztály egyszerre tartartalmazza az adott egységet, valamint permutáltjait és/vagy reverz komplementjeit (**`acg=cga=gac=cgt=gtc=tcg`**). Annak érdekében, hogy azonosítani tudjuk az adott tökéletes ismétlődéseknek megfelelő nem tökéletes ismétlődéseket (azaz, hogy kiválasszuk a tökéletes és a nem tökéletes ismétlődéseket a vizsgált genom azonos lókuszán), minden nem tökéletes ismétlődés, ami az egyes tökéletes ismétlődések helye körül azonosítottunk, kiválasztásra került. Ha több nem tökéletes ismétlődésű párja volt az adott tökéletes ismétlődésnek, azoknak az ismétlődéseknek adtunk prioritást, amelyek az adott tökéletes ismétlődéssel azonos osztályba kerültek. Ha nem volt ilyen nem tökéletes ismétlődés, az ismétlődő egységek hosszát vettük figyelembe, úgy, hogy egyiküknek a másik többszörösének kellett lennie (pl. egy tökéletes ismétlődés 6 tagú ismétlődő egységgel, párba került egy olyan nem tökéletes ismétlődéssel, amelynek 3 tagú ismétlődő egysége volt). Nagyon fontos megemlíteni**,** hogy ebben a kontextusban az egyes tökéletes és nem tökéletes ismétlődés párok ugyanabban a genomban helyezkednek el és nem próbáltunk homológ ismétlődésekre vonatkozó szisztematikus keresést végezni a rokon genomokban.

1. táblázat. A tökéletes ismétlődések (perfect repeats) és a nem tökéletes ismétlődések (imperfect repeats) kapcsolata

Tökéletes ismétlődés	**aacaacaacaac**
Nem tökéletes ismétlődés (trf)	aag**aacaacaacaac**atcaacaa
Nem tökéletes ismétlődés (int)	aag**aacaacaacaac**atcaac

A „trf" a `Tandem Repeats Finder` program által megtalált nem tökéletes ismétlődés. A `TRF` által megtalált SSR-ek eltéréseket is tartalmazhatnak (mismatch, inszerció, deléció) és az ismétlődő egység nem egész számú többszörösei is az azonosított ismétlődések közé tartoznak. Példánkban az **aac** 7,66-szor, két eltéréssel fordul elő. Az „int"a „trf"változata, melyet a részleges ismétlődések levágásával hoztunk létre. Az „int" elnevezés az angol „integer" szó rövidítése, melynek jelentése: „egész", tekintve, hogy egész számú ismétlődésekről van szó a nem egész részek elhagyásával a bal szélen.

4.3.2. Az adatok tárolása, felodolgozása

Minden adatot MySQL (Widenius *és* Axmark 2002) táblázatokban tároltunk a további analízis számára. Azonos lókuszokon található tökéletes és nem tökéletes párokat (azaz nem tökéletes ismétlődésként is azonosított tökéletes ismétlődéseket és tökéletes ismétlődésként is azonosított nem tökéletes ismétlődéseket) azonosítottunk. Ezért ezeket az ismétlődéseket két független módszerrel (is) azonosítottuk. Az SSR-eket megjelöltük/felosztottuk aszerint, hogy kódoló vagy nem kódoló régióban találhatók az NCBI annotáció CDS rekordja alapján. A rokonokban található ortológ géneket a KEGG adatbázisból- Kyoto Encyclopedia of Genes and Genomes (Kanehisha *és* Goto 2000) kerestük ki. A szekvenciákat a `CLUSTALW` programmal illesztettük. Minden további/egyéb programozást igénylő lépést saját készítésű `PERL` programokkal hajtottunk végre. A trinukleotid SSR-ek által kódolt aminosav ismétlődések azonosítása úgy történt, hogy a GenBank fájlok annotációja alapján a mikroszatelliták által lefedett fehérje szekvenciát vettük figyelembe. Csak azokat az aminosavakat vettük be az elemzésbe, amelyek ebben a fehérje szekvenciában domináltak, azaz, több mint 50%-ban voltak jelen. Ez azért volt fontos, mert még a tökéletes ismétlődések

is- az ismétlődési egységek és a kodonok esetleges egybe nem esése miatt – többféle aminosavat kódolhatnak.

4.3.3. Az ismétlődések elemzése

Annak érdekében, hogy megállapíthassuk a tökéletes és a nem tökéletes ismétlődések eloszlásai közötti különbséget és hogy összehasonlíthassuk az ismétlődések eloszlását a különböző baktérium törzsekben χ^2 kontingencia analízist végeztünk. Ez az analízis a két megfigyelt eloszlás azonosságát teszteli. A motívum/minta (pattern) szó itt az ismétlődések a teljes genom különböző mikroszatellita osztályaiban tapasztalt relatív gyakoriságát (relative abundance) jelenti. Az adatok az 5. és a 6. táblázatban, valamint az 1. kiegészítő táblázatban találhatók. Hangsúlyozzuk, hogy ebben az analízisben kapott valószínűség (probabilitás érték) értelmezése nem úgy történt, mint a hagyományos statisztikai tesztekben, hanem ez a fajta probabilitás inkább egy mutató, ami az eloszlások hasonlóságát méri (Gáspári *és mtsi*. 2007). Ezentúl amikor ezt a valószínűséget a genomom belüli tökéletes és nem tökéletes ismétlődés-párok összehasonlítására használjuk, nem várunk statisztikailag szignifikáns különbségeket, mivel az összehasonlított adatok ugyanazon szekvencia különböző értelmezéseiből származnak. Ezért bármilyen tapasztalt különbözőség biológiailag releváns eltérésre utal(hat). A teszt alkalmazása során azokat az ismétlődés osztályokat, amelyek az összes osztályra kapott teljes érték 5%-ánál kevesebb értéket kaptak, elvetettük egy iteratív közelítés alkalmazásával, melyben a legkisebb adatpontokat eltávolítottuk, míg az összes megmaradt osztályra kapott érték legalább 5%-a volt a teljes értéknek. Ez azért volt szükséges, hogy a teszt alkalmazása kellőképpen erős legyen (Townend 2004). Annak érdekében, hogy az analízist még megbízhatóbbá tegyük, és azért, hogy a tökéletes és nem tökéletes ismétlődések különböző teljes gyakoriságából fakadó vagy a különböző genomokból származó ismétlődések által okozott eltérést kiküszöböljük, minden adatot normalizáltunk a χ^2 értékek kiszámolását megelőzően. A kontingencia analízist mind a megabázisonkénti mikroszatellitaszámra, mind pedig a megabázisonkénti összmikroszatellitahosszra elvégeztük.

A számolás részletei a következők:

Az *m* kategóriák két sokaságára/hányadára tapasztalt értékek, ($O(1,1), O(2,1), ..., O(m,1), O(1,2), O(2,2)$, …, $O(m,2)$) felhasználásával a várt értékek (*E*) kifejezése:

$$E(i, j) = \frac{O(i,x)O(x, j)}{O(x,x)}$$

Ahol *O* az angol "observed" (megfigyelt) szóból, *E* pedig az angol "expected" (várt) szóból származtatható rövidítés, $O(i,x)$ és $O(x, j)$ a megfelelő sor és oszlop összegek és $O(x,x)$ az összes megfigyelés teljes összege (a normalizált hossz/megabázis adatok használata során $O(x, j) = 100$ és $O(xx) = 200$). A χ^2 érték kifejezése:

$$\chi^2 = \sum_{j=1}^{2} \sum_{i=1}^{m} \frac{[O(i, j) - E(i, j)]^2}{E(i, j)}$$

Végül a *P* valószínűséget (probabilitás) kiszámítottuk a χ^2 értékből m-1 szabadsági fok esetén. Az így kapott 0,5 alatti valószínűség értékek jelzik, hogy a két összehasonlított sokaság különbözik. Megközelítésünk nagyon hasonló a `PRIDE` módszerhez, amit fehérje szerkezet-összehasonlításra alkalmaznak (Carugo *és* Pongor 2002).

5. Eredmények

5.1. Teljes *Chlamydia* és *Escherichia coli* genom szekvenciák

Vizsgálataink kezdetekor (2001 vége) 5 *Chlamydia* teljes genom szekvenciája volt elérhető az adatbázisokból. Az 5 *Chlamydia* genom szorosan kapcsolt filogenetikai csoportot alkot. Genomonként hozzávetőlegesen 1000 gént azonosítottak. Obligát intracelluláris paraziták. Genomméretük aránylag kicsi (a *Chlamydia trachomatis* D/UW-3/CX törzsé 1,04 Mbp (megabázispár, millió bp), a *Chlamydia muridarum* Nigg törzsé 1,07 Mbp, a *Clamydophila pneumoniae* CWL029 törzsé 1,23 Mbp). Összehasonlításképpen az extracelluláris életmódú *Escherichia coli* K-12 MG1655 törzs genommérete 4,6 Mbp. 2003-ra további 2 *Chlamydia* genom szekvenciája készült el és vált elérhetővé az adatbázisokból, így a mikroszatelliták eloszlásának vizsgálatát 7 a *Chlamydiales* rendbe tartozó genomban végezhettük.

Napjainkban a *Chlamydiák* a legintenzívebben szekvenált mikroorganizmusok közé tartoznak. A dolgozat írásának időpontjában (2008. április) 13 teljes *Chlamydia* genomszekvencia szerepel az Entrez Genome adatbázisban[a]. A vizsgált 7 teljes genomszekvencián kívül a *Candidatus Protochlamydia amoebophila* UWE25, a *Chlamydia trachomatis* 434/Bu, a *Chlamydia trachomatis* A/HAR-13, a *Chlamydia trachomatis* L2b/UCH-1/proctitis, a *Chlamyophila abortus* S26/3 *és a Chlamyophila felis* Fe/C-56 genomok szekvenciája is elérhető. A mikroszatellita-eloszlás vizsgálatban (2004-ben) 4 *Escherichia coli* törzs teljes genom szekvenciáját hasonlítottuk össze a *Chlamydia* genomokkal. A dolgozat írásának időpontjában 13 teljes *Escherichia coli* genom szekvencia szerepel az Entrez Genome adatbázisban. A vizsgált 4 teljes genom szekvencián kívül az: *Escherichia coli* 536 (UPEC), az *Escherichia coli* APEC O1 (APEC)[b], az *Escherichia coli* C ATCC 8739, az *Escherichia coli* E24377A, az *Escherichia coli* HS, az *Escherichia coli* SECEC SMS-3-5 az *Escherichia coli* UTI89 (UPEC), az *Escherichia coli* K-12 DH10B és az *Escherichia coli* K-12 W3110 genom szekvenciája is elérhető (2. táblázat).

[a]http://www.ncbi.nlm.nih.gov/genomes/static/eub.html

[b]madárpatogén *E. coli* (*a*vian *p*athogenic *E*scherichia *c*oli)

2. táblázat. A teljes *Chlamydia* és *Escherichia coli* genomok adatai

A vizsgált törzs neve	Accession (GenBank adatbázis)	GN[c]	RefSeq (GenBank adatbázis)	Teljes hossz (bp)	A kódoló régiók hossza (bp)	HGT kimutatás	SSR vizsgálat
***Candidatus Protochlamydia amoebophila* UWE25**	BX908798.1	***pcu***	NC_005861.1	2414465	1979861		
***Chlamydia muridarum* Nigg**	AE002160.2	***cmu***	NC_002620.2	1072950	961248	X	X
***Chlamydia trachomatis* 434/Bu**	AM884176.1	***ctb***	NC_010287.1	1038842	914181		
***Chlamydia trachomatis* A/HAR-13**	CP000051.1	***cta***	NC_007429.1	1044459	929569		
***Chlamydia trachomatis* D/UW-3/CX**	AE001273.1	***ctr***	NC_000117.1	1042519	936164	X	X
***Chlamydia trachomatis* L2b/UCH-1 /proctitis**	AM884177.1	***ctl***	NC_010280.1	1038869	914205		
***Chlamyophila abortus* S26/3**	CR848038.1	***cab***	NC_004552.2	1144377	995608		
***Chlamydophila caviae* GPIC**	AE015925.1	***cca***	NC_003361.3	1173390	1046055		X
***Chlamyophila felis* Fe/C-56**	AP006861.1	***cfe***	NC_007899.1	1166239	1049615		
***Chlamydophila pneumoniae* AR39**	AE002161.1	***cpa***	NC_002179.2	1229853	1090813	X	X
***Chlamydophila pneumoniae* CWL029**	AE001363.1	***cpn***	NC_000922.1	1230230	1085960	X	X
***Chlamydophila pneumoniae* J138**	BA000008.3	***cpj***	NC_002491.1	1226565	1097297	X	X
***Chlamydophila pneumoniae* TW-183**	AE009440.1	***cpt***	NC_005043.1	1225935	1102622		X
***Escherichia coli* 536**	CP000247.1	***ecp***	NC_008253.1	4938920	4296860		
***Escherichia coli* APEC O1**	CP000468.1	***ecv***	NC_008563.1	5082025	4370542		
***Escherichia coli* C ATCC 8739**	CP000946.1	**-**	NC_010468.1	4746218	4081747		
***Escherichia coli* CFT073**	AE014075.1	***ecc***	NC_004431.1	5231428	4600495		X
***Escherichia coli* E24377A**	CP000800.1	***ecw***	NC_009801.1	4979619	4232676		
***Escherichia coli* HS**	CP000802.1	***ecx***	NC_009800.1	4643538	3993443		
***Escherichia coli* O157:H7 EDL933**	AE005174.2	***ece***	NC_002655.2	5528445	4820481		X
***Escherichia coli* O157:H7 Sakai**	BA000007.2	***ecs***	NC_002695.1	5498450	4819150		X
***Escherichia coli* SECEC SMS-3-5**	CP000970.1	***ecm***	NC_010498.1	5068389	4409498		
***Escherichia coli* UTI89**	CP000243.1	***eci***	NC_007946.1	5065741	4457852		
***Escherichia coli* K-12 DH10B**	CP000948.1	***ecd***	NC_010473.1	4686137	3889494		
***Escherichia coli* K-12 MG1655**	U00096.2	***eco***	NC_000913.2	4639675	4048916		X
***Escherichia coli* K-12 W3110**	AP009048.1	***ecj***	AC_000091.1	4646332	3995846		

[c]GN: a KEGG adatbázis rövidítései, a horizontális géntranszfer és a mikroszatellita vizsgálatban szereplő genomokat a táblázatban megjelöltük. A horizontális géntranszfer vizsgálathoz a szekvencia adatokat az NCBI honlapjáról töltöttük le 2001. február 19-én. A fehérje szekvenciák a GenBank formátumú fájlok annotációjából származnak, a mikroszatellita vizsgálatban szereplő GenBank formátumú szekvencia adatok az NCBI honlapjáról származnak, a letöltés dátuma: 2004. január 25. A feldolgozott genomoknál az akkor megadott hosszadatokat vettük figyelembe.

5.2. Horizontális géntranszfer események kimutatása

5.2.1. A hasonlósági keresés és a filogenetikai analízis eredményei

A filogenetikai fákat a nukleotid szekvencia illesztések alapján `TREE-PUZZLE` programmal *quartet puzzling* algoritmussal készítettük. A `BLAST` hasonlósági keresések eredményei közül kiválasztott homológ géncsoportok illesztett szekvenciáiból `PAUP*` programmal, maximális parszimónia (MP, maximum parsimony) algoritmus felhasználásával szerkesztettünk filogenetikai fákat, bootstrap analízist végeztünk.

A TREE-PUZZLE program 7-féle tartalmú és kiterjesztésű kimeneti fájlt állít elő (3. táblázat). Ezek közül a .puzzle kimeneti fájl a TREE-PUZZLE jegyzőkönyv, a .tree eredményfájl pedig a végső *quartet puzzling* fa.

3. táblázat. A TREE-PUZZLE program által szolgáltatott kimeneti fájlok

kiterjesztés	**alapértelmezett fájl név**	**a fájl tartalma**
.puzzle	outfile	TREE-PUZZLE jegyzőkönyv
.dist	outdist	ML távolságok
.tree	outtree	végleges fa/fák
.qlist	outqlist	a nem megoldott kvartettek listája
.ptorder	outptorder	az egyedi *puzzling* lépésekből származó fa topológiák listája
.pstep	outpstep	az egyedi *puzzling* lépésekből származó fa topológiák listája kronológikus sorrendben
.eps	outlm.eps	A *likelihood mapping analízis*ben készült EPS grafikus file

ML távolságok = maximum likelihood távolságok. A paraméterek becslése és a páronkénti ML távolságok legyártása filogenetikai fa szerkeztése/rekonstruálása nélkül is lehetséges. Ilyenkor a TREE-PUZZLE csak két kimeneti fájlt szolgátat: .puzzle és .dist. A TREE-PUZZLE felajánlja a *likelihood mapping analízis*t, egy a belső ágak támogatottságának vizsgálatára alkalmas módszert, anélkül, hogy elkészítné a teljes fát és grafikusan szemléltetné a szekvencia illesztés filogenetikai tartalmát. Ez a file tartalmaz egy Encapsulated Postcript formátumot (EPSF). (Schmidt *és mtsi* 2000 nyomán).

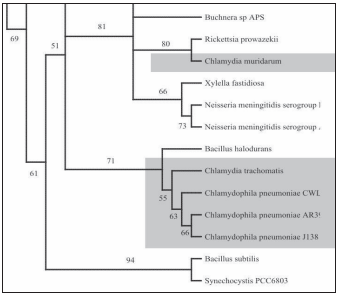
Példaként a *groEL* génnel kapcsolatos 2 eredményfájlt (groEL_2.dat.puzzle, groEL_2.dat.tree) és a statisztikai analízis egyik eredményfájlját (boostrap.log) mutatom be (1. kiegészítő ábra).

A három különböző filogenetikai módszer eredményfáinak konszenzusfáit együtt vizsgálva választottuk ki azokat, ahol a *Chlamydia* csoportjában tapasztalt monofíliatörés (azaz legalább 1 az 5 *Chlamydia* genomból valamelyik másik leszármazási ággal került egy klaszterbe (Stanhope *és mtsi*. 2001) vagy valamelyik másik faj csatlakozott a *Chlamydia* genushoz) több módszerrel is kimutatható volt. Ezenfelül a HGT eseménynek másik kritériuma a minimum 50%-os támogatottság (PUZZLE Support Value) volt. Példaként a putatív szukcinil-transzferázt kódoló *sucB* génre szerkesztett filogenetikai fákat mutatjuk be

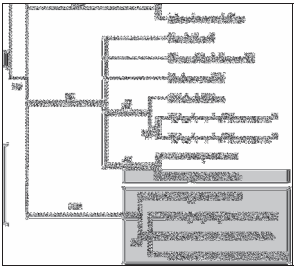
(7. ábra). A *sucB* feltételezett donorja a *Rickettsia prowazekii*, a *C. muridarum* genomba való bekerülését mindhárom filogenetikai módszerrel készült konszenzusfa támogatta.

Tizennégy HGT eseményt sikerült megbízhatóan kimutatni. Az élőlények mindhárom doménje képviselteti magát a feltételezett donorok között: 5 esetben a Baktériumok (Bacteria), 4-ben az Eukarióták (Eukaryota), 3-ban az Archaeák (Archaea) közül kerültek ki a feltételezett donorok (8. ábra és 4. táblázat). Érdekes módon a *trpC*, *ychB*, and *CPn0608* géneknek növény HGT donorja, a *pfrA* gén közeli rokonságot mutat a gombákkal, azonban egyetlen állati eredetű HGT érintett gént sem találtunk.

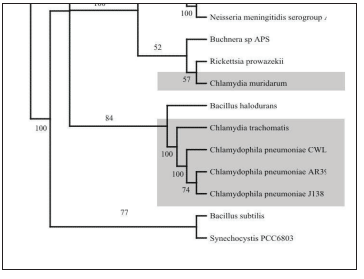
A)



B)

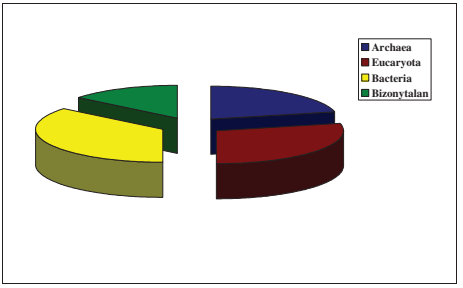


C)

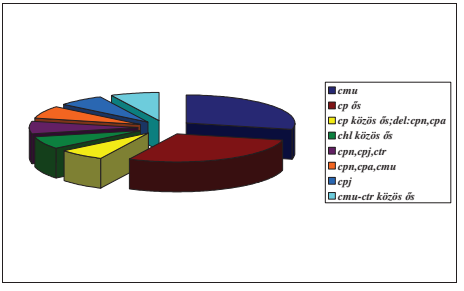


7. ábra A *sucB* génre és homológjaira készült fák

A: A `TREE-PUZZLE` programmal készült *quartet puzzling* fa, **B:** Maximális parszimónia módszerrel (Maximum Parsimony) készült fa, **C:** A szomszéd-összevonó módszerrel (Neighbor Joining) készült fa, az **A** ábrarészen a támogatottságot kifejező (`PUZZLE` Support Value) értékek, a **B** és **C** ábrarészek fáinak ágain a bootstrap értékek láthatók. A *Chlamydia* genomokat szürke színnel emeltük ki.



8. ábra. A detektált HGT események eloszlása feltételezett donorok szerint



9. ábra. A detektált HGT események eloszlása befogadó törzsek szerint

cmu*: *Chlamydia muridarum Nigg, ***cp ős*: *Chlamydophila pneumoniae*** ős, ***cp közös ős*: *Chlamydophila pneumoniae*** közös ős, ***del***: deléció, ***cpn*: *Chlamydophila pneumoniae*** CWL029, ***cpa*: *Chlamydophila pneumoniae*** AR39, ***chl közös ős*: *Chlamydia*** közös ős, ***cpj***: ***Chlamydophila pneumoniae*** J138, ***ctr*: *Chlamydia trachomatis*** D/UW-3/CX.

4. táblázat. A *Chlamydiák*ban detektált HGT eseményekre vonatkozó adatok

Génnév (GenBank-GeneID)	**Feltételezett funkció**	**A HGT esemény feltételezett donora (Domén)**	**A befogadó *Chlamydia* törzs**	***Quartet puzzling* fa (tám.)**[d]	**NJ[c]fa (bootstrap érték)**	**MP[c]fa (bootstrap érték)**	**a GC tartalom eltérése**
TC0062, oppD (1245591)	ABC transzporter ATP-áz	*Thermoplasma acidophilum (Archaea)*	*Chlamydia muridarum*	tám. (93)	igen (100)	igen (100)	nem (0,23%)
TC0255, xerC (1246425)	integráz/rekombináz *XerD*	*Rickettsia prowazekii (Eubacteria)*	*Chlamydia muridarum*	tám. (93)	igen (50)	nem (<50)	nem (0,97%)
TC0325, sucB (1246368)	szukcinil transzferáz	*Rickettsia prowazekii (Eubacteria)*	*Chlamydia muridarum*	tám. (81)	igen (100)	igen (100)	nem (0,28%)
TC0339 (1246383)	ABC transzporter ATP-kötő fehérje	*Treponema pallidum (Eubacteria)*	*Chlamydia muridarum*	tám. (75)	igen (88)	igen (79)	nem (0,18%)
CPj0576_1 (919336)	peptid-lánc 2-es release faktor	*Bizonytalan*	*Chlamydophila pneumoniae* J138	tám. (88)	igen (100)	igen (100)	nem (4,14%)
CP0565 (962796)	peptid ABC transzporter ATP-kötő fehérje	*Bizonytalan*	*Chlamydophila pneumoniae közös ős*	tám. (94)	igen (57)	igen (64)	nem (*cpn*:0,41%, c*pa*:0,38%, *cpj*:0,43%)
TC0061, dppF (1245590) és *CT689 (884490)*	peptid ABC transzporter ATP-kötő fehérje	*Archaea*	*Chlamydia muridarum-Chlamydia trachomatis közös ős*	tám. (89)	igen (98)	igen (95)	nem (*ctr*:0,64%, *cmu*:0,23%)
CPn0777, groEL (895077)	Hsp 60 (hősokk fehérje)	*Desulfobacterium dehalogenans (Eubacteria)*	*Chlamydophila pneumoniae közös ős*	tám. (98)	igen (98)	igen (98)	nem (*cpn*:1,54%, *cpa*:1,63%, *cpj*:1,58%)

Génnév (GenBank-GeneID)	**Feltételezett funkció**	**A HGT esemény feltételezett donora (Domén)**	**A befogadó *Chlamydia* törzs**	***Quartet puzzling* fa (tám.)[e]**	**NJ[d]fa (bootstrap érték)**	**MP[e]fa (bootstrap érték)**	**a GC tartalom eltérése**
CPn0113, pfrA (895574)	peptid-lánc releasing faktor	*Fungi* (*Eukaryota*)	*Chlamydophila pneumoniae* CWL029, AR39 *Chlamydia muridarum*	tám. (85)	igen (91)	igen (93)	nem (*cpn*:0,78%, *cpa*:0,76%, *cmu*:3,53%)
CPn0403, yceC (895037)	prediktált pszeudo-uridin szintetáz enzim család	*Thermotoga maritima* (*Eubacteria*)	*Chlamydophila pneumoniae* CWL029, J138 *Chlamydia trachomatis*	tám. (96)	igen (66)	igen (50)	nem (*cpn*:2,16%, *cpa*:2,14%, *ctr*:0,27%)
CPn0486 (895220)	hipotetikus prolin permeáz	*Pyrococcus abyssi* (*Archaeaa*)	*Chlamydophila pneumoniae közös ős*	tám. (100)	igen (100)	igen (100)	nem (*cpn*:3,07%, *cpa*:3,80%, *cpj*:3,05%)
CPn0608 (895475)	uridin 5'-monofoszfát szintetáz	*Nicotiana plumbaginifolia* (*Eukaryota*)	*Chlamydophila pneumoniae közös ős*	tám. (71)	igen (88)	igen (68)	nem (*cpn*:0,20%, *cpj*:0,10%, *cpa*:0,10%)
CPj0954, ychB (919718)		*Mentha piperita* (*Eukaryota*)	*Chlamydophila pneumoniae* J138, *Chlamydia muridarum*, *C. trachomatis közös ős*	nem tám. (<50)	igen (90)	nem (<50)	nem (*cpj*:1,59%, *ctr*:1,41%, *cmu*:0,56%)
CT327, trpC (884791)	foszforibozil-antranilát izomeráz	*Arabidopsis thaliana* (*Eukaryota*)	*Chlamydia közös ős*	tám. (98)	igen (98)	igen (75)	nem (*ctr*:0,38%)

[d]Támogatottság (PUZZLE Support Value)

[e]Szomszéd összevonó módszer (Neighbor Joining)

[f]Maximális parszimónia módszer (Maximum Parsimony)

A HGT események többségét a *Chlamydia muridarum*ban és a *C. pneumoniae* törzsekben detektáltuk, míg a *C. trachomatis* ritkábban vett részt horizontális géntranszferben. Ez azt sugallja, hogy a laterális géntranszfer különböző rátával jelenik meg ezekben a közel rokon genomokban. Említésre méltó, hogy 4 esetben a *Chlamydia muridarum* egyedül volt befogadó genom, további 7 esetben feltételezzük, hogy a HGT az ág közös ősében történhetett, ezért jelent meg a gén az ág valamennyi leszármazottjában. A horizontálisan transzferált gének több esetben is ABC transzporter gének, az ATP-kötő kazetta (ATP-binding casette) által kódolt fehérjék gyakran hozhatók kapcsolatba a baktériumok körében gyakran kialakuló drogrezisztenciával (Poelarends *és mtsi*. 2002; Burnie *és mtsi*. 2002). Várakozásunkkal ellentétesen nem detektáltunk HGT-t az ún. plaszticitás zónában (plasticity zone, Read *és mtsi*. 2000), mely a *C. muridarum* és a *C. pneumoniae* genomokban a prediktált terminációs helyekhez közel található (4. táblázat és 9. ábra).

5.3. A mikroszatellita-eloszlás vizsgálat eredményei

5.3.1. A kiválasztott genomok mikroszatellita eloszlásainak összefoglalása

Általánosan elmondható, hogy a vizsgált genomokban a tri- és a hexanukleotid ismétlődések túlsúlyát tapasztaltuk, sokkal kevesebb mono-, di-, tetra- és pentanukleotid ismétlődés volt detektálható. Ez az eredmény konzisztens a vizsgált genomok nagy részét alkotó kódoló szekvenciákkal is. Mivel csak a 3 és sokszorosainak megfelelő egységek hossza nem okoz frameshift típusú mutációt a kódoló régiókban történő kiterjedés (expanzió) során. Valójában az ismétlődések döntő többsége a kódoló régiókban található (lásd 1. kiegészítő táblázat). A legtöbb azonosított tökéletes ismétlődésnek van „nem tökéletes ismétlődés párja" (azaz a tökéletes ismétlődést nem tökéletes ismétlődésként is azonosítottuk). Különbségek akkor jelentek meg, amikor a nem tökéletes ismétlődés más ismétlődő egységgel volt azonosítható. Tipikusan az $\mathbf{a_n x}$ egységek ilyenek (pl. **aaag** és **aaaaat**), melyeket a TRF nem tökéletes ismétlődésű polyA szakaszként azonosított. A tökéletes ismétlődések hossza legtöbbször 12 nukleotid (nt, pl. 3 tagú egységek), ami a megengedett minimálisan detektálható SSR hosszúság analízisünkben. A leghosszabb átlagos ismétlődés hosszt az *E. coli* O157:H7 és az *E. coli* O157:H7 EDL933 törzsekben figyeltük meg, ahol az **agagcc** ismétlődések hossza általában 34, ill. 46 nt feletti egyenként. Bár 1-6 egységű SSR-eket azonosítottunk, annak megfelelően, hogy a tanulmányozott genomok mindegyike halmozottan tartalmaz fehérje szekvenciákat, ebben az értekezésben főként a trinukleotid ismétlődésekre fókuszáltam. A nem tökéleteseként is azonosított tökéletes és a tökéleteseként is azonosított nem tökéletes trinukleotid ismétlődések (itt ez szigorúan csak ugyanabba az ismétlődés osztályba tartozó ismétlődéseket jelent) megabázisonkénti halmozott hossz adatait kiszámoltuk a 4 *E. coli* és a 7 *Chlamydia* genomban (5. és 6. táblázat, 2. kiegészítő ábra).

5.3.2. A trinukleotid ismétlődések áttekintése a vizsgált genomokban

Annak ellenére, hogy az azonosított ismétlődések száma általában véve kicsi (egyes ismétlődés osztályokat ötnél kevesebb tag alkot, sőt egyes esetekben egyetlen 12 nt hosszúságú kópia található a teljes genomban), néhány tendenciát (trendet) egyértelműen lehetett azonosítani.

Megállapítható, hogy a vizsgálatba bevont *Chlamydia* genomok általában nagyobb SSR sűrűséget mutatnak, mint a vizsgált *Escherichia coli* törzsek (5. és 6. táblázat). A

megfigyelt különbség több mint kétszeres a tökéletes ismétlődéseket tekintve (a *C. caviae*-ben ~345 bp /Mbp, ezzel szemben az *E. coli* O157:H7 EDL33 esetében ~142 bp/Mbp). Tekintetbe véve az azonosított ismétlődéseket, nincsenek kiugró különbségek az általános ismétlődés gyakoriságok között az egyes törzsekben, de az *E. coli* genomok kb. 1,5-ször több ismétlődés típust tartalmaznak, mint a a *Chlamydia* genomok (lásd 1. kiegészítő táblázat). Az *E. coli* törzsek előnyben részesítik az **acc**, **agc**, **atc** és a **ccg** trinukleotid ismétlődéseket. Ahogyan az várható volt, a két O157:H7 törzs szoros evolúciós rokonságukból kifolyólag, nagyon hasonló mikroszatellita eloszlást mutat. A legnagyobb nem tökéletes ismétlődés- tökéletes ismétlődés arányt is ebben a két genomban találtuk az **aac** trinukleotid ismétlődések esetében (»1,75). Ez azt jelenti, hogy a nem tökéletes ismétlődésű mikroszatelliták átlagosan 1,75-ször hosszabbak, mint a nekik megfelelő tökéletes ismétlődés-változatok (azaz tökéleteseként is azonosított nem tökéletes ismétlődések). Mindegyik *Chlamydia* genom gazdag **aag** trinukleotid ismétlődésekben, bár a mikroszatellita-eloszlás kevésbé homogén, mint az *E. coli* törzsekben. A *Chlamydia muridarum* törzsben az **agc** trinukleotid ismétlődés kimagaslóan a leggyakoribb ismétlődés (6. táblázat, 10. ábra és 2. kiegészítő ábra), majdnem olyan gyakori, mint a C. *caviae-ben* megfigyelt maximális **aag** trinukleotid ismétlődés, míg a **ccg** trinukleotid ismétlődést csak a *C. caviae* törzsben lehetett megfigyelni. A három *C. pneumoniae* törzs hasonló mikroszatellita eloszlást mutat. A legnagyobb nem tökéletes ismétlődés/ tökéletes ismétlődés arányt a *C. pneumoniae* törzsek **aac** (>2,7) és a *C. muridarum* törzsek **agc** (1,9) mikroszatellita osztályában találtuk.

A tökéletes trinukleotid ismétlődések gyakorisága az összes (azaz kódoló és nem kódoló) szekvenciára a *Chlamydiák*ban durván fele, az *E. coli* törzsekben hozzávetőlegesen harmada az eukariótákban találtaknak hasonló azonosító kritériumok használatával (Tóth *és mtsi.* 2000). Annak ellenére, hogy ez a különbség nagy mértékben emelkedik (2-es vagy 3-as szorzó faktor taxontól függően), amikor baktérium adatokat eukarióta kódoló szekvenciákkal hasonlítunk össze, ez azt jelzi, hogy a a mi vizsgálati kritériumaink szerint a teljes trinukleotid ismétlődés sűrűség (denzitás) teljes egészében nem különbözik drasztikusan a prokarióta és az eukarióta genomokban. Az azonos GC tartalmú, de különböző ismétlődés gyakoriságok (pl. **aac**, **aag**, **agc** és **acg**) azt sugallják, hogy az általános ismétlődés eloszlást szelektív evolúciós erők befolyásolják, nem pedig nemspecifikus sztochasztikus folyamatok eredménye.

5. táblázat. A nem tökéletesként is azonosított tökéletes trinukleotid ismétlődések és a tökéletesként is azonosított nem tökéletes trinukleotid ismétlődések teljes hossza megabázisonként a 4 teljes *Escherichia coli* genomban (bp/Mbp)

	eco[g]		***ecc***		***ecs***		***ece***	
	tökéletes	nem tökéletes	tökéletes	nem tökéletes	tökéletes	nem tökéletes	tökéletes	nem tökéletes
aac	12,93	19,18	6,88	7,46	9,28	16,19	9,22	16,10
aag	12,93	14,87	13,76	16,06	8,73	10,37	8,68	10,31
aat	10,35	10,35	6,88	6,88	10,91	10,91	10,85	10,85
acc	28,45	29,10	38,99	43,01	19,64	24,55	19,54	24,42
acg	12,93	12,93	6,88	6,88	6,55	8,18	6,51	8,14
act	2,59	2,59	2,29	2,29	2,18	2,18	2,17	2,17
agc	28,45	37,51	25,23	32,69	26,19	35,47	28,22	37,44
atc	31,04	42,46	32,11	45,69	32,74	42,38	32,56	42,69
ccg	32,33	42,46	38,42	50,85	24,55	28,92	24,42	28,76
ÖSSZES	**172,00**	**211,45**	**171,44**	**211,81**	**140,77**	**179,15**	**142,17**	**180,88**

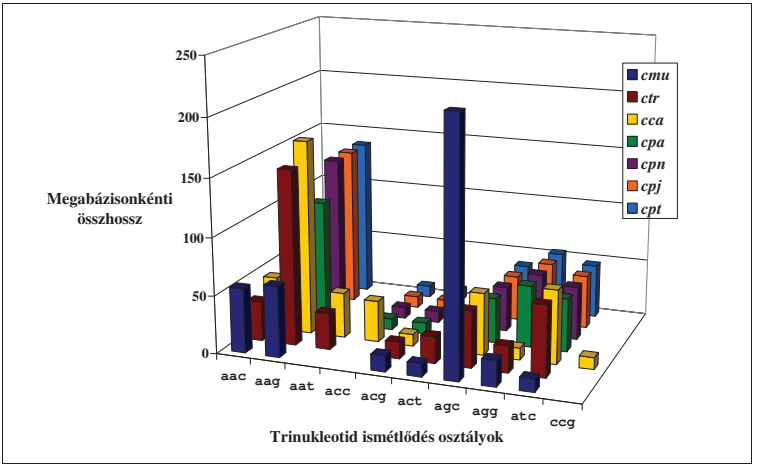
Csak az azonos mikroszatellita osztályba tartozó ismétlődéseket összegeztük.[g]A három betűs rövidítések megfelelnek a KEGG adatbázisnak genom azonosítójának (GN), lásd 2. táblázat.

6. táblázat. A nem tökéletesként is azonosított tökéletes trinukleotid ismétlődések és a tökéletesként is azonosított nem tökéletes trinukleotid ismétlődések teljes hossza megabázisonként a 11 teljes *Chlamydia* genomban (bp/Mbp)

	cmu[h]		***ctr***		***cca***		***cpa***		***cpn***		***cpj***		***cpt***	
	tökéletes	nem tökéletes	tökéletes	nem tökéletes	tökéletes	nem tökéletes	tökéletes	nem tökéletes	tökéletes	nem tökéletes	tökéletes	nem tökéletes	tökéletes	nem tökéletes
aac	47,53	55,92	34,53	34,53	40,91	46,02	9,76	26,83	9,75	26,82	9,78	26,90	9,79	26,92
aag	55,92	61,51	103,59	151,56	132,95	168,74	82,94	107,33	102,42	137,37	102,73	137,78	102,78	137,85
aat	0,00	0,00	25,90	31,65	10,23	39,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
acc	0,00	0,00	0,00	0,00	30,68	35,79	9,76	9,76	9,75	9,75	9,78	9,78	9,79	9,79
acg	13,98	13,98	14,39	14,39	10,23	10,23	9,76	9,76	9,75	9,75	9,78	9,78	9,79	9,79
act	11,18	11,18	23,02	23,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
agc	114,64	218,09	48,92	48,92	43,46	53,69	29,27	39,03	29,26	39,02	29,35	39,13	29,36	39,15
agg	22,37	22,37	23,02	23,02	10,23	10,23	51,23	53,66	51,21	53,65	51,36	53,81	51,39	53,84
atc	11,18	11,18	34,53	61,39	56,25	63,92	39,03	46,35	39,02	46,33	39,13	46,47	39,15	46,49
ccg	0,00	0,00	0,00	0,00	10,23	10,23	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	276,80	**394,23**	**307,90**	**388,48**	**345,17**	**438,05**	**231,75**	**292,72**	**251,16**	**322,69**	**251,91**	**323,65**	**252,05**	**323,83**

Csak az azonos mikroszatellita osztályba tartozó ismétlődéseket összegeztük. [h]A három betűs rövidítések megfelelnek a KEGG adatbázisnak genom azonosítójának (GN), lásd 2. táblázat.

Néhány érdekes különbséget tapasztaltunk, amikor a trinukleotid ismétlődések által kódolt kódolt aminosav szekvenciákat hasonlítottuk össze a *Chlamdiák*ban és az *E. coli* törzsekben (2. kiegészítő táblázat). A tökéletes **aag** ismétlődések főleg/jobbára glutaminsavat kódolnak a *Chlamydiák*ban, az *E. coli-ban* viszont fenilalanint. Bár az előbbi a leggyakoribb aminosav az összes törzsben, a nem tökéletes ismétlődésekben, a fenilalanin kétségtelenül jelen van az *E. coli* genomokban. Mind a tökéletes, mind a nem tökéletes **agc** ismétlődések esetében a kódoló szekvenciákban világosan látszik a különbség, az alanin a *Chlamydiák*ban, a leucin az *E. coli* törzsekben leggyakoribb.



10. ábra. A tökéletesként is azonosított nem tökéletes trinukleotid ismétlődések eloszlás mintázata a *Chlamydia* törzsekben

A törzsek rövidítései a következők: ***cmu:*** *Chlamydia muridarum* Nigg, ***ctr*****:** *Chlamydia trachomatis* D/UW-3/CX, ***cca:*** *Chlamydophila caviae* GPIC, ***cpa:*** *Chlamydophila pneumoniae* AR39, ***cpn:*** *Chlamydophila pneumoniae* CWL029, ***cpj:*** *Chlamydophila pneumoniae* J138, ***cpt*****:** *Chlamydophila pneumoniae* TW-183, a KEGG adatbázis rövidítéseinek átvételével (ld. 2. táblázat, az összes vizsgált törzs eloszlás mintázatát lásd 2. kiegészítő ábra).

5.3.3. A tökéletes és a nem tökéletes trinukleotid ismétlődések összehasonlítása

Az összes tökéletes és nem tökéletes trinukleotid ismétlődés genomeloszlását a 3. kiegészítő ábra mutatja. Ahogy ez várható volt, a nem tökéletes ismétlődések sokkal több helyen voltak detektálhatók a genomban, mint a tökéletes ismétlődések. Nyilvánvaló, hogy a

sokkal kisebb genomméret bizonyos ismétlődés osztályok (pl. **aat**, **act**, **ccg**) hiányát idézi elő a legtöbb *Chlamydia* törzsben, annak ellenére, hogy a normalizált teljes SSR gyakoriság nagyobb a *Chlamydiales*-en belül, mint az *E. coli* esetében. Annak érdekében, hogy a tökéletes és a nem tökéletes ismétlődések összehasonlíthatók legyenek, kontingencia vizsgálatot végeztünk (mikroszatellitaszám megabázisonként és összmikroszatellitahossz megabázisonként) a nem tökéletes ismétlődésként is azonosított tökéletes és a tökéletes ismétlődésként is azonosított nem tökéletes trinukleotid ismétlődésekből származó adatsokaságon (5. és 6. táblázat). Az 1 körüli P értékek az jelzik, hogy a két összehasonlított eloszlás nagyon hasonló, míg a kisebb probabilitás érték eltéréseket/különbözőségeket jeleznek. Ez azt jelenti, hogy a nem tökéletes ismétlődés osztályok relatív gyakorisága nem követi a megfelelő tökéletes „párok" - vagyis nem tökéletesként is azonosított tökéletes ismétlődés osztályok - relatív gyakoriságát (7. táblázat). Míg a tökéletes/nem tökéletes eloszlások nagyon hasonlóak a legtöbb vizsgált genomban, nagyon különböznek a *Chlamydia muridarum*ban. Az ismétlődés adatokat közelebbről megvizsgálva, látható, hogy ezt **agc** ismétlődések teljes hosszában való eltérések okozzák. A nem tökéletes **agc** szakaszok teljes hossza közel kétszerese a tökéletes ismétlődésekének, míg a többi ismétlődés osztály egyáltalán nem vagy csak finom eltéréseket mutat a tökéletes és a nem tökéletes ismétlődések hosszában megadva (10. ábra és 2. kiegészítő ábra).

7. táblázat. A nem tökéletes ismétlődésként is azonosított tökéletes trinukleotid ismétlődések és a tökéletes ismétlődésként is azonosított nem tökéletes trinukleotid ismétlődések egybeesésének valószínűsége az összes vizsgált törzs minden régiójában

Törzs[i]	***cmu***	***ctr***	***cca***	***cpa***	***cpn***	***cpj***	***cpt***	***ecc***	***eco***	***ecs***	***ece***
hossz per megabázis	**0,30**	0,89	1,00	0,94	0,92	0,92	0.92	0,98	0,99	0,99	0,99
Szám per megabázis	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

A 0,5 alatti valószínűséget vastaggal szedtük. [i]A három betűs rövidítések megfelelnek a KEGG adatbázisnak genom azonosítójának (GN), lásd 2. táblázat. A nagyobb értékek szorosabb hasonlóságot mutatnak a 0-tól 1-ig tartó skálán (a részleteket lásd. 5.2.3. fejezet). A részletes adatok a 3. kiegészítő táblázatban találhatók.

Meg kell jegyezni, hogy ez a törzs nem tökéletes **aag** ismétlődéseket is tartalmaz nagy gyakorisággal, de ezek többségének nincs ugyanolyan ismétlődés osztályba tartozó megfelelő tökéletes ismétlődés párja. Ezért nem nagyon valószínű, hogy ezek tökéletes **aag** szakaszokból származnak. Ez a péda illusztrálja, hogy milyen fontos az analízis során, hogy a

függetlenül azonosított ismétlődéseket mind tökéletesként, mind nem tökéletesként megvizsgáljuk.

5.3.4. A különböző baktériumokban megfigyelt mikroszatellita-eloszlás összehasonlítása

A nem tökéletes ismétlődésként is azonosított tökéletes és a tökéletes ismétlődésként is azonosított nem tökéletes ismétlődések eloszlását is összehasonlítottuk a bakteriális genomok között (8. táblázat és 3. kiegészítő táblázat). Mind a *Chlamydia* mind az *E. coli* genomok esetében általában az egyes tagok hasonlósága csoporton belül nagyobb, mint a két taxonómiai csoport tagjai között, ami gyakorlatilag zérus. Az egyetlen említésre méltó kivétel a *C. muridarum*, amely nagyobb hasonlóságot mutat 3 *E. coli* genomhoz (is). Ez csak a megabázisonkénti mikroszatellita hosszaknál figyelhető meg, a megabázisonkénti mikroszatellitaszámok esetében azonban nem. A *Chlamydiales* renden belül a változatos *C. pneumoniae* törzsek nagymértékben hasonlítanak egymásra ismétlődés gyakoriságokban kifejezve. Ezek a megfigyelések különböző SSR preferenciákat jeleznek a két vizsgált filogenetikai csoportban. Kis valószínűségeket kaptunk akkor is, amikor a különböző *Chlamydia* fajokat hasonlítottuk össze.

8. táblázat. Az ismétlődés eloszlások hasonlóságága a különböző genomokban az eloszlások egybeesésének valószínűségével mérve

A). Minden régió

	cmu[j]	*ctr*	*cca*	*cpa*	*cpn*	*cpj*	*cpt*	*ecc*	*eco*	*ecs*	*ece*
cmu		0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.045	0.554	0.412
cmu		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ctr	0.000		0.627	0.033	0.064	0.064	0.064	0.000	0.000	0.000	0.000
ctr	0.001		0.730	0.043	0.097	0.097	0.097	0.000	0.000	0.000	0.000
cca	0.000	0.992		0.257	0.162	0.162	0.162	0.000	0.000	0.000	0.000
cca	0.000	0.730		0.206	0.158	0.158	0.158	0.000	0.000	0.000	0.000
cpa	0.000	0.150	0.960		0.908	0.908	0.908	0.000	0.000	0.000	0.000
cpa	0.000	0.043	0.206		0.895	0.895	0.895	0.000	0.000	0.000	0.000
cpn	0.000	0.256	0.900	0.946		1.000	1.000	0.000	0.000	0.000	0.000
cpn	0.000	0.097	0.158	0.895		1.000	1.000	0.000	0.000	0.000	0.000
cpj	0.000	0.256	0.900	0.946	1.000		1.000	0.000	0.000	0.000	0.000
cpj	0.000	0.097	0.158	0.895	1.000		1.000	0.000	0.000	0.000	0.000
cpt	0.000	0.256	0.900	0.946	1.000	1.000		0.000	0.000	0.000	0.000
cpt	0.000	0.097	0.158	0.895	1.000	1.000		0.000	0.000	0.000	0.000
ecc	0.083	0.000	0.000	0.000	0.000	0.000	0.000		0.890	0.362	0.297
ecc	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.890	0.369	0.304
eco	0.144	0.000	0.000	0.000	0.000	0.000	0.000	0.833		0.957	0.945
eco	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.861		0.953	0.940
ecs	0.265	0.000	0.000	0.000	0.000	0.000	0.000	0.442	0.955		1.000
ecs	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.301	0.953		1.000
ece	0.331	0.000	0.000	0.000	0.000	0.000	0.000	0.366	0.935	1.000	
ece	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.249	0.940	1.000	

B) Kódoló régiók

	cmu[j]	*ctr*	*cca*	*cpa*	*cpn*	*cpj*	*cpt*	*ecc*	*eco*	*ecs*	*ece*
cmu		0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.033	0.177	0.152
cmu		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ctr	0.000		0.479	0.033	0.050	0.050	0.050	0.000	0.000	0.000	0.000
ctr	0.000		0.549	0.043	0.070	0.070	0.070	0.000	0.000	0.000	0.000
cca	0.000	0.629		0.257	0.213	0.213	0.213	0.000	0.000	0.000	0.000
cca	0.000	0.549		0.206	0.190	0.190	0.190	0.000	0.000	0.000	0.000
cpa	0.000	0.125	0.960		0.985	0.985	0.985	0.000	0.000	0.000	0.000
cpa	0.000	0.043	0.206		0.983	0.983	0.983	0.000	0.000	0.000	0.000
cpn	0.000	0.186	0.970	0.997		1.000	1.000	0.000	0.000	0.000	0.000
cpn	0.000	0.070	0.190	0.983		1.000	1.000	0.000	0.000	0.000	0.000
cpj	0.000	0.186	0.970	0.997	1.000		1.000	0.000	0.000	0.000	0.000
cpj	0.000	0.070	0.190	0.983	1.000		1.000	0.000	0.000	0.000	0.000
cpt	0.000	0.186	0.970	0.997	1.000	1.000		0.000	0.000	0.000	0.000
cpt	0.000	0.070	0.190	0.983	1.000	1.000		0.000	0.000	0.000	0.000
ecc	0.218	0.000	0.000	0.000	0.000	0.000	0.000		0.829	0.494	0.497
ecc	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.827	0.510	0.507
eco	0.544	0.000	0.000	0.000	0.000	0.000	0.000	0.780		0.781	0.848
eco	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.867		0.772	0.836
ecs	0.890	0.000	0.000	0.000	0.000	0.000	0.000	0.497	0.904		1.000
ecs	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.451	0.772		1.000
ece	0.941	0.000	0.000	0.000	0.000	0.000	0.000	0.498	0.954	1.000	
ece	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.475	0.836	1.000	

A tökéletes ismétlődéseket a mátrix felső háromszögeiben, a nem tökéletes ismétlődéseket pedig a mátrix alsó háromszögeiben hasonlítottuk össze. Minden törzs első sora (szürke háttérrel) a megabázisonkénti teljes hosszra vonatkozó statisztikát tartalmazza, minden törzs második sora (fehér háttérrel) pedig a megabázisonkénti ismétlődések számát tartalmazza. A nagyobb értékek szorosabb hasonlóságot mutatnak a 0-tól 1-ig tartó skálán (a részleteket lásd. 5.2.3. fejezet). A számok az eloszlások egybeesésének valószínűségét jelentik, tehát azt, hogy mi a valószínűsége annak, hogy a

két minta ugyanabból az eloszlásból származik. A részletes adatok a 3. kiegészítő táblázatban találhatók. [1]A baktérium törzsek nevének rövid formáját a KEGG adatbázisnak megfelelően használtuk, lásd 2. táblázat.

5.3.5. A legnagyobb SSR tartalmú gének

A legnagyobb SSR tartalmú géneket azonosítottuk és vizsgáltuk az analízisben szereplő összes (11) genomban. Csak azokat az ismétlődéseket vettük figyelembe, amelyeket mind tökéletes, mind pedig nem tökéletesként is azonosítottunk, de a trinukleotid ismétlődések analízisével ellentétben a különböző ismétlődés egységgel rendelkező SSR-eket is bevettük az analízisbe (pl. egy hosszú nem tökéletes trinukleotid ismétlődésnek több tökéletes hexanukleotid ismétlődés szakasz lehet párja, ebben az esetben a nem tökéletes ismétlődést nem vetettük el). A *Chlamydiák*ban számos azonosított gén vagy polimorfikus külső membránfehérjéket (Pmps) vagy hipotetikus fehérjéket kódol. A genomonkénti 10-10 legnagyobb mikroszatellita-tartalmú gén részletes adatai a 4. és az 5. kiegészítő táblázatban találhatók.

Chlamydia törzsek

A legnagyobb összmikroszatellita tartalmú gén (a tökéletes ismétlődések hossza ~100 bp) a *C. muridarum*-ban található: *TC0438.* Ez a gén egy tapadási faktort (adherence factor) kódol, habár a többi *Chlamydiá*ban található ortólog gének nem tartalmaznak SSR-eket. A *C. muridarum dnaE* génje (*TC0832*) is viszonylag nagy (68 bp) összmikroszatelliata tartalommal rendelkezik. Ezenkívül a polimorfikus külső membrán fehérjéket (Pmp-ket) kódoló gének is nagy összmikroszatellita tartalommal bírnak: ez a *pmp_15*, a *pmp_18* és a *pmp_20* esetén ~50 bp, a *pmp_21* esetén pedig 64 bp (6. kiegészítő táblázat). A *C. muridarum* 50S riboszómális L7/L12 fehérjét kódoló *rplL* génje (*TC0590*) is a legnagyobb SSR tartalmú gének közé tartozik, mivel a benne található nem tökéletes ismétlődésű mikroszatelliták összhossza 63 bp. A *C. muridarum*-ban összesen 13 olyan gént azonosítottunk, amelyek nem tökéletes és további 9 gént, melyek tökéletes ismétlődésű **agc** ismétlődéseket tartalmaznak. A következő 9 gén egyaránt tartalmaz tökéletes és nem tökéletes **agc** ismétlődéseket: *TC0107* hipotetikus fehérje, *TC0181*, *glgA*, glikogén szintáz, *TC0283*, phoH foszfát-éhezés-indukált fehérjéhez kapcsolódó fehérje, *TC0371*, *infB*, transzlációs iniciációs faktor, *TC0431*, MAC/perforin családba tartozó protein/fehérje, *TC0440*, foszfolipáz D családba tartozó fehérje, *TC0590*, *rplL*, 50S riboszómális fehérje L7/L12, *TC0867*, hipotetikus fehérje és *TC090*, hipotetikus fehérje. A fennmaradó gének, a *TC0433*, foszfolipáz D családba tartozó proteint/fehérjét, a

TC0486, hipotetikus fehérjét, a *TC0500* és a *TC0908*, ugyancsak hipotetikus fehérjéket kódoló gének csak nem tökéletes **agc** ismétlődéseket tartalmaznak (4. és 5. kiegészítő táblázat).

***E. coli* törzsek**

A vizsgált 4 *Escherichia coli* törzsben 2 kimagaslóan nagy mikroszatellita tartalmú gént találtunk: az *ftsK* és a *tolA* gént. Az *ftsK* génben a nem tökéletes ismétlődések összhossza ~ 300 bp, míg a *tolA* gén az *E. coli* CFT073 és az *E. coli* K-12 törzsek egy 171 bp hosszúságú, az *E.coli* O157:H7 és az *E. coli* O157:H7 EDL933 törzsek pedig egy 291 bp hosszúságú nem tökéletes ismétlődést tartalmaznak (4. és 5. kiegészítő táblázat).

Funkcionális kategóriák

A nagy tökéletes ismétlődés-tartalmú génjeinek funkcionális kategóriáit tekintve (a KEGG adatbázis génkatalógusát használtuk) látható, hogy az anyagcserével kapcsolatos gének messze a leggyakoribbak 11 vizsgált genomban. A 110 megfigyelt génből azonban csak 42-őt katalogizáltak a KEGG adatbázisban. Ebből a 42 génből 29 az Anyagcsere osztályba tartozik. Ezen a csoporton belül a legtöbb gén (9) a szénhidrát anyagcserével kapcsolatos. Ezek közül 4 az aminosav anyagcserével, míg 3 gén a környezeti információ feldolgozással, 1 gén az energiával kapcsolatos anyagcserével áll kapcsolatban, a fennmaradó 1 gén pedig a keményítő és szacharóz anyagcserében vesz részt. További 5 gén sorolható az aminosav anyagcserével kapcsolatos csoportba. Ezek a gének a genetikai információ feldolgozással is kapcsolatosak: 4 gén az enzim családokkal (peptidázok), 4 gén a glükán bioszintézissel és anyagcserével, 2 gén a lipid anyagcserével hozható kapcsolatba; 2 gén tartozik a nukleotid (purin és pirimidin) anyagcsere alkategóriába (az utóbbi 2 gén a replikáció és reparáció alkategóriával is kapcsolatban áll), további 2 gén a kofaktorok és vitaminok anyagcseréjéhez kapcsolódik, a fennmaradó 1 génnek pedig az energiával kapcsolatos anyagcserében (valamint a környezeti információ feldolgozásban) van szerepe. Mindössze 11 gén vesz részt információ kezeléssel kapcsolatos folyamatokban (6 gén sorolható a környezeti információ feldolgozás, 5 pedig a genetikai információ feldolgozás alkategóriába). Végül 2 a Humán betegségek fő kategóriába sorolható gént azonosítottunk (9. táblázat és 5. kiegészítő táblázat).

9. táblázat. A tíz leghosszabb ismétlődést tartalmazó gén eloszlása a vizsgált 11 genomban a funkcionális kategóriák szerint, a KEGG adatbázis kategóriáinak megfelelően

KEGG Funkcionális kategória	A tökéletes ismétlődéseket tartalmazó gének száma	A nem tökéletes ismétlődéseket tartalmató gének száma
Anyagcsere Kategória	**29**	**8**
Szénhidrát anyagcsere Alkategória	9	2
Szénhidrát anyagcsere és Aminosav anyagcsere Alkategóriák	4	1
Szénhidrát anyagcsere és Környezeti Információ Feldolgozás Alkategóriák	3	1
Szénhidrát anyagcsere és Energia anyagcsere Alkategóriák	1	
Szénhidrát anyagcsere és Keményítő és szacharóz anyagcsere Alkategóriák	1	
Aminosav anyagcsere és Genetikai információ feldolgozás Alkategóriák	5	1
Ezim családok és Genetikai információ feldolgozás Alkategóriák	4	
Glükán bioszintézis és anyagcsere Alkategória	4	
Lipidanyagcsere Alkategória	2	2
Nukleotid (purin és pirimidin) anyagcsere és Replikáció és reparáció Alkategóriák	2	
Kofaktorok és vitaminok anyagcseréje Alkategória	2	
Energia anyagcsere Alkategória	1	1
Környezeti információ feldolgozás Kategória	**6**	**15**
Környezeti Információ Feldolgozás és Sejtszintű folyamatok Alkategóriák	5	
Környezeti Információ Alkategória	1	
Környezeti információ feldolgozás és Sejtszintű folyamatok (sejtosztódás) Alkategóriák		4
Membránranszport/ABC transzporterek Alkategória		4
Membránranszport/Pórus[k] ioncsatornák Alkategória		4
Membránranszport/Egyéb ion-csatolt transzporterek Alkategória		2
Membránranszport/Foszfotranszferáz rendszer Alkategória		1
Genetikai információ feldolgozás Kategória	**5**	**4**
Genetikai információ feldolgozás és Replikáció és reparáció Alkategóriák	3	
Genetikai információ feldolgozás és Transzláció Alkategóriák	3	3
Genetikai információ feldolgozás Alkategória	1	1
Humán betegségek Kategória	**2**	**2**
A KEGG adatbázis által nem besorolt	**68**	**81**

A kép egy kicsit más, ha a nagy nem tökéletes ismétlődéseket vesszük alapul. Ekkor a környezeti információ feldolgozás kategóriába tartozó gének alkotják a legnagyobb funkcionális csoportot, az anyagcserével kapcsolatos géneket a 2. helyre szorítva. Ezeket a megfigyeléseket másképp megfogalmazva azt is mondhatjuk, hogy a környezeti információ feldolgozásban szerepet játszó gének nem tökéletes ismétlődéseiben található tökéletes szakaszok relatív hossza kisebb, mint az anyagcserével kapcsolatos kódoló szekvenciákban. Meg kell jegyeznünk azonban, hogy a legtöbb gént még nem sorolták be a KEGG génkatalógusába, ezért ezeket nem tudtuk bevenni ebbe az elemzésbe.

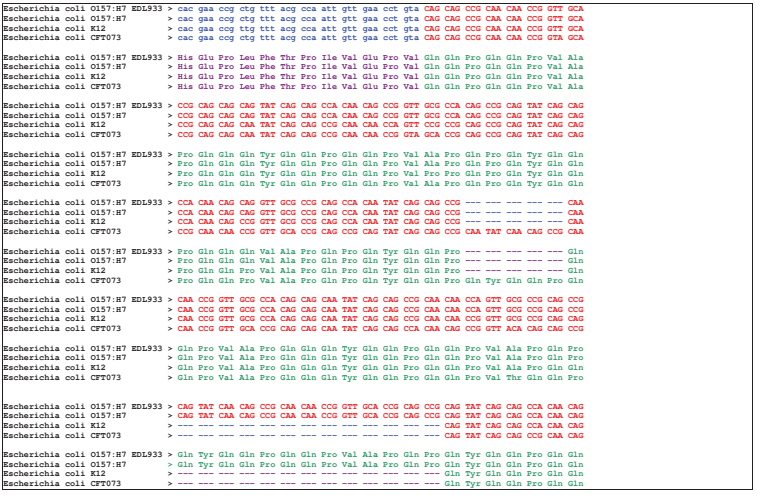
[k]Sok Gram negatív baktérium szekréciós rendszerének része, mely mind a belső mind a külső membrán felé nyitott.

5.3.6. Génspecifikus összehasonlítás, összehasonlítás a gének szintjén

Kiválasztottunk néhány olyan funkcionálisan jól jellemezhető gént, melyek legalább 2 vizsgált genomban tartalmaznak mikroszatellitákat, azért, hogy kihangsúlyozzuk a tökéletes és a nem tökéletes ismétlődések egyidejű analízisének fontosságát/jelentőségét az evolúciós folyamatok értelmezésében.

A sejtmembrán fehérjéket kódoló *tolA* gén

A leghosszabb nem tökéletes ismétlődések a vizsgálatban szereplő 4 *Escherichia coli* törzs sejtmembrán fehérjéket kódoló *tolA* génjében találhatók. Ez a fehérje nincs jelen a *Chlamydiák*ban. A `Tandem Repeats Finder` (`TRF`) egy **`agc`** trinukleotid egységet tartalmazó nem tökéletes ismétlődésnek sorolja be ezt a mikroszatellitát, ugyanakkor nincs megfelelő (azaz nem tökéletesként is azonosított) tökéletes ismétlődés ugyanebben a mikroszatellita osztályban. Rövid hexanukleotid ismétlődés szakaszok azonban találhatók ebben a régióban. A 6 bp hosszú ismételt egység szekvenciája mindig **`agc`** alszekvenciát (pl. **`aagcag`**) tartalmaz. Két törzsben a nem tökéletes ismétlődés hossza 171 bp, míg a két O157:H7 törzs esetében az ismétlés hossza 291 bp. Az ebből a 4 *E. coli* törzsből származó *tolA* gén szekvenciák nagyon hasonlóak, két izoformjuk van. Az O157:H7 törzsek a rövidebbiket tartalmazzák/hordozzák. Bár a hosszú **`agc`** ismétlődés egy 27 tagból álló inszerciós maradék körül található, a hosszabb SSR változatok/variánsok a rövidebb izoformokban találhatók. A szekvenciák behatóbb tanulmányozásakor látszik, hogy egy nagyon hasonló ismétlődés régiót azonosítottunk a két esetben (11. ábra és 4. kiegészítő ábra). Mivel a *tolA* gének eltérő aminosav gyakoriságokra szelektálódtak más források (Rooney 2003) szerint és nem találhatók tökéletes ismétlődések a megfelelő tökéletes ismétlődés osztályokban (nem tökéletesként is azonosított tökéletes ismétlődés osztályok) ebben a génben, feltételezzük, hogy ennek a lókusznak a nem tökéletes ismétlődése nem egy ősi tökéletes ismétlődésből származik, hanem a kodon gyakoriságra irányuló szelekció eredménye(ként jöhetett létre), bár egy ilyen ismétlődés-kiterjesztés forgatókönyvet nem lehet teljesen kizárni.



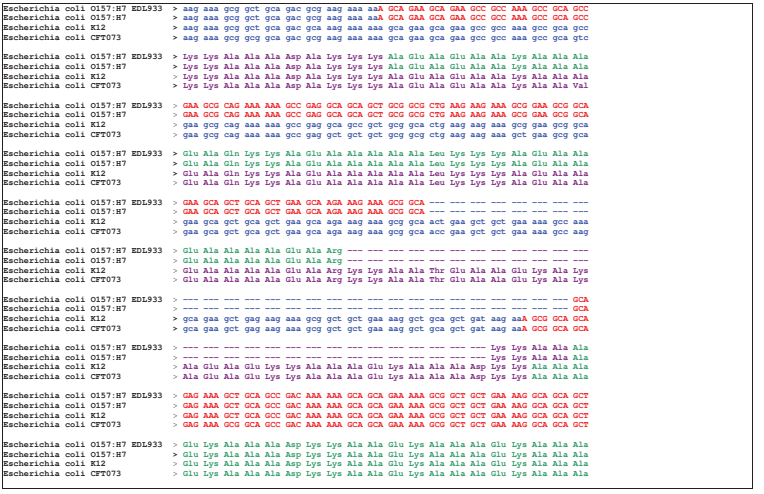
11. ábra. Részlet a *tolA* gén **agc** mikroszatellita régiójának illesztéséből a 4 vizsgált *Escherichia coli* törzsben

Az illesztett nukleotid szekvenciákban piros nagybetűvel emeltük ki a mikroszatellitát tartalmazó szakaszt, a mikroszatellitát nem tartalmazó többi szakaszt pedig kék kisbetűvel. Az illesztett aminosav szekvenciákban a mikroszatellitának megfelelő szakaszt zöld szín, a többi szakaszt pedig lila szín jelzi, a teljes illesztést lásd 4. kiegészítő ábra.

A sejtosztódással kapcsolatos fehérjéket kódoló ftsK gén

Az FtsK egyike a DNS átvitelben (transzferben) résztvevő bakteriális fehérjéknek (Errington *és mtsi.* 2001), valamint az az eszenciális bakteriális ATP-áz, ami helyes kromoszóma szegregációjáért felelős a sejtosztódás során. Különböző bakteriofágok és eukarióta kétszálú DNS vírusok DNS-csomagoló/bepakoló ATP-ázai is az FtsK-HerA szupercsaládba tartoznak (Iyer *és mtsi.* 2004). Az FtsK fehérje, amely összekapcsolja a kromoszóma szegregációt és a sejtosztódást az *E. coli*ban ATP-függő DNA-transzlokáznak látszik. A XerCD-függő rekombinációt aktiválja a kromoszóma dimérek monomerekké való átalakításával azáltal, hogy a szál vágás rendjét átkapcsolja a rekombináz alegységekkel (Donachie 2002; Aussel *és mtsi.* 2002; Massey *és mtsi.* 2004).

Az FtsK fehérjét kódoló gén, az *ftsK* a mikroszatellitákban leginkább gazdag gének közé tartozik mind a tökéletes, mind pedig a nem tökéletes ismétlődésű mikroszatellitákat tekintve a vizsgált *Escherichia coli* törzsekben, ám ugyanez nem mondható el a vizsgált *Chlamydiák* esetében. A *tolA* génekhez hasonlóan az *ftsK* génben egy hosszú **agc** ismétlődés van túlsúlyban/uralkodik, melynek csak különböző hexanukleotid szekvenciák tökéletes ismétlődései feletetehetők meg (12. ábra és 4. kiegészítő ábra).



12. ábra. Részlet az *ftsK* gén **agc** mikroszatellita régiójának illesztéséből a 4 vizsgált *Escherichia coli* törzsben

Az illesztett nukleotid szekvenciákban piros nagybetűvel emeltük ki a mikroszatellitát tartalmazó szakaszt, a mikroszatellitát nem tartalmazó többi szakaszt pedig kék kisbetűvel. Az illesztett aminosav szekvenciákban a mikroszatellitának megfelelő szakaszt zöld szín, a többi szakaszt pedig lila szín jelzi, a teljes illesztést lásd 4. kiegészítő. ábra.

Az 50S nagy riboszómális alegységet felépítő L7/L12 fehérjét kódoló gének

Az *rplL* gén, mely az 50S nagy riboszómális alegységet felépítő L7/L12 fehérjét kódolja a vizsgált *E. coli* törzsek mindegyikében tartalmaz egy 45 bp hosszú nem tökéletes **agc** ismétlődést, melyben egy 12 bp hosszú tökéletes ismétlődés-szakasz is megtalálható.

Mivel a nukleotid szekvencia majdnem teljesen konzervált, az ismétlődések megegyeznek ebben a 4 *E. coli* törzsben. Ez a DNS szekvencia ennek az alaninben gazdag fehérje szekvenciának feleltethető meg: AAAAVAVAAGPVEAA. A *Chlamydiák*ban is tartalmaznak **agc** vagy **aacagc** mikroszatellita osztályba tartozó nem tökéletes ismétlődéseket ezek a gének, de ezeknek nem feleltethetők meg tökéletes ismétlődések. A nagy ismétlődés-tartalmú gének további adatai (baktérium törzs név, genom, NCBI-GI és ACCESSION, ismétlődéshossz) az 5. kiegészítő táblázatban találhatók.

6. Az eredmények értékelése

6.1. Horizontális géntranszfer kimutatása

Horizontális géntranszfert főleg a *C. muridarum* és a *C. pneumoniae*. Törzsekben detektáltunk, a *Chlamydia trachomatis*ban a HGT ritka eseménynek tekinthető, megállapítható tehát, hogy a HGT különböző rátájú ezekben a szoros filogenetikai kapcsolatú törzsekben. A *Chlamydiák* nem mutattak HGT preferenciát szignifikánsan egyik taxonómiai csoport felé sem, az élőlények mindhárom doménje (Eukaryota, Archaea, Bacteria) képviselteti magát a feltételezett donorok között. A kapott adatok világosan mutatják, hogy a HGT-t nem gátolják az egyes specifikus csoportok, hanem a HGT átlépi a taxon határokat.

Sikerült egy konszenzus alapú módszert találnunk, mellyel megbízhatóan ki tudtuk mutatni HGT eseményeket. és tudomásunk szerint ez az első szisztematikus filogenetikai alapú megközelítés, mely prokariótákban a szerzett gének megbízható kimutatását tűzte ki célul. Megállapíthatjuk, hogy bár a *Chlamydiák* a magasabb rendű eukarióták obligát intracelluláris parazitái, s ezáltal feltehetőleg jobban elzártak a HGT-től, mint a szabadon élő fajok, eredményeink megmutatják, hogy diverzifikációjukban szerepet kapott idegen szekvenciák genomjaikba való felvétele is.

Analízisünk eredményei jelentősek lehetnek abból a szempontból is, hogy képet kaphatunk ezen orvosbiológiai szempontból fontos patogén szervezetek rövid távú evolúciójáról, és ezzel együtt hozzájárulhatunk e törzsek diverzifikációjának pontosabb megértéséhez.

6.2. Egyéb genomevolúciós események kimutatása

A részletes filogenetikai analízis mintegy „melléktermékeként" sikerült detektálnunk egy olyan teljes génkészletet melynek tagjaiban deléció, duplikáció vagy genomátrendeződés következett be a *Chlamydia* vonal két fajra és öt törzsre történt szétválása óta tartó evolúciós időszak alatt. Ezen genomevolúciós események többsége ismert volt a *Chlamydia* szakirodalomban is (Read *és mtsi*. 2000). Kiemelkedően fontosnak tartom, hogy közel 200 duplikációt sikerült kimutatni, valamint figyelemre méltó az is, hogy a *Chlamydophila pneumoniae* törzsekben jelentős duplikációs aktivitás figyelhető meg, a *Chlamydophila*

pneumoniae J138 törzsben egy 41 tagú duplikációs csoportot találtunk. Ezek a gének putatív polimorf membránfehérjéket (Pmp-ket) kódolnak, melyek a gazdaszervezet immunrendszere elleni változatos felszíni antigénként játszhatnak szerepet. Ráadásul, ezek az erősen repetitív génkópiák is klaszterekbe rendeződnek a kromoszóma mentén, mely jelzi a tandem génduplikációs események szerepét evolúciós történetükben. Nem meglepő, hogy a Pmp-k a mikroszatellita vizsgálatok eredményeinek értékelése során is fókuszba kerültek (lásd 6.5. fejezet). Az egyéb genomevolúciós események kimutatásával kapcsolatos vizsgálatok és eredmények bővebben a kutatással kapcsolatos másik doktori értekezésben kerültek bemutatásra (Ortutay 2003).

6.3. Kodon-eloszlás vizsgálat

A kodonhasználati eltérésekkel kapcsolatos megközelítés alapja az, hogy a közelmúltban HGT-vel szerzett gének inkább a donorra mint a gazda genomra jellemző karakterisztikus kodon összetételt mutatnak. A szerzett géneknek át kell esniük egy kodon összetételükre vonatkozó változáson, hogy alkalmazkodjanak, „igazodjanak/illesztődjenek" (amelirotate) a genom maradék részéhez. Széleskörűen elfogadott nézet, miszerint a túlexpresszált gének egy adaptált kodonkészletet használatára evolválódtak a transzkripciós és transzlációs gépezet nagy hatékonysággal való működtetése érdekében (Sharp *és* Li 1986). A fehérjék aminosav maradékait kódoló génekben a kodonok kiválasztása/megválasztása összhangban van a sejtekben nagy bőségben jelenlevő rokon transzfer RNS molekulák mennyiségével, ezért a riboszóma elongációs ciklust nem gátolja a szükséges adott tRNS-hozzáférhetősége. Ez az -egyszerűsített- elképzelés azzal a ténnyel együtt, hogy a kodon preferenciák taxonra jellemzőek/taxonspecifikusak, hozták létre a „illeszkedés" elméletet, például azt, hogy az idegen baktériumfajból HGT útján szerzett géneknek egy kodonkészlet-változáson kell átesniük azért, hogy fehérje szinten is kifejeződhessenek. Minthogy ehhez evolúciós skálán mérve időre van szükség, a genom azon a génjeinek, amelyek a közelmúlban lezajlott HGT eseményből származnak detektálhatónak kell lenniük kizárólag a rájuk jellemző kodon összetétel alapján, ha a már a gazda genomhoz „igazított" génekhez (feltételezhetően a genomban található gének többsége) viszonyítjuk őket. Tekintve, hogy napjainkban csak a kodon összetételen alapuló HGT detektáló módszereket lehet alkalmazni nagy léptékű (genomi szintű) vizsgálatokhoz, szükségessé vált, ezen közelítések megbízhatóságának ellenőrzése.

Munkacsoportunk vizsgálta a kodon adaptációs index (*C*odon *A*daptation *I*ndex, CAI, Sharp *és* Li 1987) és egy általunk bevezetett mérőszám, az egyesített távolság (CDT, *C*umulative *D*istance) HGT-re vonatkozó diagnosztikus jelentőségét is.

A CAI ($0 \leq CAI \leq 1$), a kodonhasználatra vonatkozó relatív alkalmazkodást fejezi ki, minél nagyobb, annál jobban „simul" az adott gén a genomba. A közelmúlban lezajlott HGT eseményből származó génekről az feltételezzük, hogy még nem volt idejük alkalmazkodni, illeszkedni a genomhoz, ezért feltehetően viszonylag kis CAI értékekkel rendelkeznek. Az egyesített távolság (CDT) az adott baktérium kodonhasználatától való euklidészi távolságból és a GC tartalom eltérésekből tevődik össze. Minél alacsonyabb az értéke, annál jobban „simul" az adott gén a genomba, ezért a magas CDT jelzi a HGT esemény(eke)t.

A fenti meggondolásoknak megfelelően azt feltételeztük, hogy a HGT-ből származó géneknek olyan CAI/CDT értékekkel kell rendelkezniük, melyek alacsony szintű (az átlagtól eltérő) génexpressziót jeleznek. Továbbá azt, hogy egy adott genomba beilleszkedettként detektált/azonosított géneknek a többi vizsgált genom többségébe is be kellett illeszkedniük. Nem vártuk, hogy egy organizmus nagy mértékben használt génjeinek homológjai éppen az ebbe a vizsgálatba bevett baktériumok körében (a közelmúlban) lezajlott HGT-ből származzanak. Minden génre kiszámítottuk a CAI-t, ahogyan azt Sharp *és* Li (1987) leírták, a COGs adatbázisban[a] (http://www. ncbi.nlm.nih.gov/cgi-bin/COG) található nagy mértékben expresszálódó génkészlet felhasználásával. A CDT-t a gén szekvenciájának következő patramétereiből számoltuk: nukleotid gyakoriság, trinukleotid gyakoriság, kodon gyakoriság, százalékos GC-tartalom eltérés az első és a harmadik kodon pozíciókban (ezeket az értékeket a genomra jellemző átlagtól való euklidészi távolság kiszámolásával kaptuk meg), valamint a gén és a gén előtti és utáni 500-500 bázispárnyi „ablak"-ra jellemző százalékos GC-tartalom-eltérést.

[a]A COGs adatbázis legújabb, 2003-as verziója már eukariótákat is tartalmaz (Tatusov *és mtsi*. 2003, http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=12969510).

Mivel a különböző szervezetek CAI és CDT értékeit közvetlenül nem lehetett összehasonlítani, ezért bevezettünk két értéket, a CAIrank és CDTrank értékeket, melyek az adott gén relatív pozíciójára jellemzőek figyelembe véve és összehasonlítva a gén CAI/CDT értékét a gazdaszervezet többi génjével. A CAIrank és a CDTrank értékeket a következőképpen számítottuk ki: a CAI és a CDT értékeket csoportosítottuk/sorba rendeztük (CAI-t csökkenő, a CDT-t növekvő értékek szerint), majd minden egyes gén az elrendezésben elfoglalt pozícióját elosztottuk a szervezet összes génjének számával. Az eredményül kapott értéket százalékos értékké alakítottuk (azaz 100-zal szoroztuk). Ezért mind a CAIrank, mind a CDTrank értékek magas/nagy értéke (közel 100%) jelzi a genomhoz jól alkalmazkodott géneket és az alacsony/kis értékek (közel 0%-hoz) tartoznak a HGT –vel nemrég szerzett génekhez.

Eredményeink alapján elmondható, hogy a kiválasztott 27 baktérium törzs esetében a CAI és a CDT csak nagyon óvatosan és korlátozott esetekben használható a HGT megbízható kimutatására. A *Chlamydia* génekre készített filogenetikai fák, csakúgy mint a BLAST hasonlósági keresések azt jelzik, hogy ezek a mérőszámok csak kevés esetben konzisztensek (egyeznek meg) az általunk várt értékekkel. Az az előzetes feltételezésünk, miszerint a rokon bakteriális géneknek hasonló CAI és/vagy CDT értékekkel (vagy *CAIrank* és/vagy *CDTrank* százalékos értekekkel) kell rendelkezniük, az esetek többségében nem teljesült. Nem vártuk, a HGT ilyen nagy hányadban való megjelenését az összes bakteriális génben. Észrevételünk, miszerint sem a CAI sem a CDT nem használható a HGT megbízható kimutatására nem ássa alá magát az „amelioration" elméletet. A javasolt jelenségek kipróbálása/tesztelése több specifikus vizsgálatot igényelne. A mi vizsgálódásunk két okból is korlátozott volt, egyrészt az ismert baktériumokak csupán egy kis csoportját vizsgáltuk, másrészt egy még kisebb génkészletet vettünk górcső alá, ezért nem is tudtunk általános következtetéseket levonni (a szerző nem publikált eredménye).

Megállapítottuk, hogy a kodonhasználat és kodon - eloszlás vizsgálat eredményei egyetlen esetben sem támogatták sem a hasonlósági keresés, sem a filogenetikai analízis eredményeit. ez összevág a szakirodalomban megjelent és közzétett más szerzők tapasztalataival (Koski *és mtsi*. 2001). A kodonhasználati eltérések mérőszámok HGT-re vonatkoztatott gyakorlati alkalmazhatóságának határait azonban sikerült kimutatni. Véleményünk szerint ezek a mérőszámok kevésbé alkalmasak a HGT megbízható kimutatására, mert a sok egyéb faktor, amely befolyásol(hat)ja a genom szerkezetét nem vonható bele ezekbe az egyszerű paraméterekbe. A kodon összetételről, (azaz a ritka vagy épp ellenkezőleg a gyakori/nagy mennyiségben jelenlevő kodonok) kimutatták, hogy befolyásolja

a riboszómális forgalmat és a kotranszlációs „feltekeredés”-t (folding) (Komar *és mtsi.* 1999), valamint azt, hogy egy adott génben is változó lehet. Ezért a HGT nagy biztonsággal való azonosítására alkalmas módszerként csak a kiválasztott gének filogenetikai analízise jöhetett szóba.

6.4. HGT kitekintés

Chlamydia genom vizsgálati eredményeink közlése (Ortutay *és mtsi*. 2003) után a szakirodalomban megjelent HGT-vel kapcsolatos kutatási eredményeket a következőképpen lehet összefoglalni:

Születtek olyan javaslatok is (Chen *és mtsi*. 2004), miszerint az intergénikus, nem transzlált szekvenciákból és nem fehérje kódoló szekvenciákból kellene becsülni a genomszintű kodonhasználati eltérést. Taoka *és mtsi*. (2004) az *Escehrichia coli* HGT-vel kapcsolatban felvetették, hogy a HGT-vel a genomba kerülő gének többsége nem expresszálódik, tehát nincs fehérje termék. Ez felveti, hogy mégis hogyan konzerválódhattak ezek a gének a bakteriális genomban. Feltételezhetően ezek közül néhány pszeudogén, melyek híján vannak azoknak a struktúrális elemeknek, melyek a transzkripcióhoz és/vagy a transzlációhoz szükségesek, esetleg olyan gének lehetnek, melyeknek nincsenek fehérje átiratai, de funkcionális RNS termelésére szolgálnak. Transzkriptom analízist (DNA microarray) végeztek és megállapították, hogy a nemrég érkezett gének többsége feltételezhetően még nem adekvát a baktérium sejt transzlációs gépezete számára. Az az érv, miszerint a HGT gének specifikus környezeti körülmények között (pl. stressz idején, ami fenyegeti a baktérium sejt túlélését) expresszálódnak. Ezzel szemben 54 stresszel kapcsolatos *Escherichia coli* gén közül mindössze 7-et kódol HGT gén.

Griffiths *és* Gupta (2006) szabadon élő *Actinobacteria* donor és *Chlamydia* akceptor közötti laterális géntranszferről számolt be. Eredményeinkkel összhangban Xia (2007) is rámutatott, hogy bár többen használták a CAI-t laterális géntranszfer detektálására, azonban a CAI pontossága és érzékenységek egy ilyen típusú vizsgálat során még értékelésre szorul. Mindazonáltal 2007-ben elérhetővé vált egy felhasználóbarát internetes szolgáltatás, a CodonO (Angelotti *és mtsi*. 2007) a genomok közötti és a genombeli kodonhasználati eltérés analízisére.

A korai módszerek, atipikus GC tartalom, kodonhasználati minták alapján való HGT prediktálás után a filogenetikai fák rekonstruálását tartották legjobb elérhető módszernek a HGT megjelenésének és irányának kimutatására, ennek azonban a legnagyobb hátránya, hogy sok programozást igényel és nagyon időigényes. 2005-2006-ban az ún. genomi aláírás (genomic signature) modellek széles választékát javasolták, változó hossz és kodon pozíció nukleotid mintáit felhasználva. Ezeket a modelleket egyenként és különböző kombinációkban is analizálták, csúszó ablakokat, Bayes osztályozók (Bayes classifier), Markov modellek és támogató vektor (support vector) módszer felhasználásával (Dalevi *és mtsi*. 2006; Waack *és mtsi*. 2006; Saunders *és mtsi*. 2005; Tsirigos *és mtsi*. 2005). Ezeknek a módszereknek az egyik korlátja azonban az, hogy valószínűsítik egy adott gén atipikus voltát, de nem szolgálnak információval származását illetően. Ennek az információnak a megszerzése és a pozitív jelöltek valós voltának igazolása egy következő, filogentikai megerősítő lépéssel történhet (Podell *és* Gaasterland 2007). 2007-ben egy új módszert (DarkHorse algoritmus) mutattak be a HGT gyors, genom szintű prediktálására, a kvantitatív, reprodukálható, programozást nem igénylő módszer kombinálható a genomi aláírás és/vagy filogenetikai faépítő eljárásokkal a pontosság és a hatékonyság növelése érdekében, alkalmazhatóságát bakteriális, archaeális és eukarióta példákkal demonstrálták (Podell *és* Gaasterland 2007). Ezen módszerek megjelenésével és alkalmazásával párhuzamosan azonban még mindig számos cikk jelenik meg, melyekben szekvencia hasonlóság keresésen (főleg `BLAST`) alapuló módszereket használnak a gén jelenlét/hiány megállapítására, ami a fajokon belüli laterális géntranszferrel történő génszerzés szisztematikus túlbecslését eredményezi (Zhaxybayeva *és mtsi*. 2007).

A HGT szerepének megítélése Lawrence *és* Hendrickson (2003) megfogalmazásában: Woese (2002) posztulálta, hogy a korai mikrobiális evolúcióban a HGT elég heves lehetett, de most kisebb szerepet játszik, a „darwini küszöbön" való átkelés után, noha az érvek impozánsak, miszerint a HGT mai szerepe különbözik az ősi leszármazási vonalakban játszott szerepétől, még mindig világos, hogy a HGT egy hatásos folyamat lehet a mikrobiális diverzifikációban. Kérdés marad hogyan lehet a HGT hatását mennyiségileg meghatározni a leszármazási vonalakban és az érdekes génekben, valamint hogyan lehet ezeket az adatokat holisztikus értelemben integrálni, a géncsere hogyan közvetít evolúciós változ(tat)ást. Koonin (2003) szerint a kulcskérdés az lesz, ha egyszer a HGT fő evolúciós hatását bebizonyítják, miért fixálódtak a horizontálisan transzferálódott gének a mikrobiális populációban? Általános válasz, hogy a fixálódást a darwini szelekció hajtja, azaz a szerzett gének meghaladják a befogadó rátermettségét (fitness-ét). Mindazonáltal nagy rejtély az, hogy a horizontálisan transzferált gének előnyt biztosíthatnak az őket megtartó sejteknek, bár olyan

donorból származnak, melynek evolúciós története igen távol esik a recipiensétől és ezért rosszul adaptálódik a működéshez az utóbbiban. A válasz viszonylag nyilvánvaló az „önző operonok" felvételének esetében, melyek új anyagcse kapacitással ruházzák fel a befogadót (Lawrence *és* Hendrickson 2003). Ezzel szemben a xenológ (ortológ) gén elmozdulás (ami talán még általánosabb formája lehet a HGT-nek) szelektív előnyét még erősebben meg kell fontolni.

A HGT kutatás érdekes aspektusa a biológiai hálózatokkal való kapcsolat. Annak ellenére, hogy a HGT formálja a bakteriális genomokat, a legtöbb nagy analízis figyelmen kívül hagyja a biológiai hálózatokra tett hatását (Pál *és mtsi*. 2005a). Taoka *és mtsi.* (2004) felvetésével kapcsolatban, miszerint a HGT gének környezet-specificitása magyarázhatja, hogy sokuknak miért nincs fehérje átirata laboratóriumi körülmények között Pál *és mtsi*. 2005a azt fogalmazták meg, hogy a hálózatok evolúcióját főként az új környezethez való adaptáció hajtja és nem a változatlan, állandó környezetben való opimalizálás, azaz optimum megtalálása. A HGT genomokat formáló hatását *Escherichia coli* anyagcsere hálózatokon (metabolic networks) vizsgálták, GC3 (GC-tartalom a 3. kodon pozícióban) és a STRING[b] (von Mering *és mtsi*. 2005) adatbázis felhasználásával, valamint génszerzés és génvesztés legtakarékosabb evolúciós forgatókönyvének (the most parsimonious evolutionary scenario, Snel *és mtsi*. 2002; Mirkin *és mtsi.* 2003; Boussau *és mtsi*. 2004) felállításával. Megállapították, hogy egy fehérjét kódoló gén HGT-vel történő sikeres felvétele egy anyagcsere hálózatba erősen függ a vele fiziológialilag kapcsolt fehérjék genombeli jelenlététől (Pál *és mtsi*. 2005b). A HGT gének a hálózat szélébe integrálódnak, míg a hálózat centrális részei evolúciós szempontból stabilak. A fiziológialiag összekapcsolódó reakciókat kódoló gének gyakran együtt transzferálódnak, tipikusan operonokban. Így a bakteriális anyagcsere hálózatoknak a megváltozott környezetre adott válasza a perifériku reakciók direkt felvételével való evolválódás (Pál *és mtsi*. 2005a). Fontos azonban megjegyezni, hogy még az alaposan tanulmányozott szervezetek genomjai is tartalmaznak ismeretlen funkciójú géneket (pl az *E. coli* esetében a gének 20%-a ilyen). Ennek eredményeképpen azok az anyagcsere hálózatok, amelyek kizárólag a genomi és a biokémiai bizonyítékokon alapulnak gyakran sok hálózati hézagot (gap-et) tartalmaznak, és a hálózati becslések nagymértékben függnek az elérhető adatoktól, különösen a fiziológiai vonatkozásúaktól (Reed *és mtsi.* 2006).

[b]Az integrálódott és szervezetek között transzferált ismert és feltételezett (prediktált) fehérje-fehérje asszociációkat tartalmazó adatbázis, melynek fejlesztett verzióját 2007-ben publikálták (von Mering *és mtsi*. 2007).

2005-ben a Nature Reviews Microbiology folyóirat egyik kiadásában egy egész fejezetben a HGT került középpontba. Gogarten *és* Townsend (2005) így foglalják össze a HGT-vel kapcsolatos információinkat:

Úgy tűnik, hogy a gének minden funkcionális kategóriája alkalmas HGT-re, még az rRNS operon és az olyan törzs (phylum) meghatározó jellemző, mint a fotoszintetikus apparátus is. De nem minden gén egyformán „vándorló". Néhánynak világosan nagyobb hajlandósága van a transzferre, mint másoknak és nem minden szervezetcsoportban tapasztalható ugyanolyan mértékű HGT. Nagymértékű géncsere történik a kromoszómális és a nem kromoszómális gének között. A legtöbb ilyen transzfer közel neutrális a befogadó számára, bizonyos körülmények között néhány ezek közül emelheti a fág vagy vírus fitnessét. A nemrégiben transzferálódott génkészletben kevés olyan gén van, amely növeli a befogadó rátermettségét (fitnessét). Ha ezekkel a transzferekkel olyan gének, géncsoportok kerülnek be a genomba, amelyek nagy azonnali előnyt jelentenek a gazdaszervezetnek, akkor gyorsan elterjedve fixálódhatnak. Ezt a folyamatot a szakirodalom szelektív söprésnek (az ango. „selective sweep" kifejezésből) nevezi. Ezen fixálódott transzferált gének azok, amelyek általánosan kimutathatók az összehasonlító molekuláris filogenetikai módszerekkel.

Annak ellenére, hogy a HGT-vel szerzett gének szekvenciális jellemzői alapján nem lehet közelebbről meghatározni, hogy melyik működőképes, a megtartott hányad becsülhető, ha tanulmányozzuk a HGT gének előfordulását a leszármazási vonal története során. A genomika végső célja, hogy megmagyarázza/értelmezze a genom minden egyes nukleotidjának biológiai szerepét és fenotípusos következményeit. Adott a legtöbb bakteriális leszármazási vonalban a gének masszív turnoverje, mely az idegen DNS szerzésén és a meglévő szekvenciák elvesztésén keresztül valósul meg, bármely adott genom fog tartalmazni „hulladékot" (debris-t), melyet az előbbi két dinamika hoz létre. Ezen helyek és régiók felvázolásának központi célnak kell lennie, ha a genomok átfogó megértését szeretnénk elérni (Ochman *és* Davalos 2006).

Várhatóan a HGT még sokáig lesz az evolúcióbiológia egyik központi kérdése.

6.5. A mikroszatellita vizsgálatok

A mikroszatellitákkal kapcsolatos vizsgálataink eredményei megmutatták, hogy a mikroszatelliták kellő bőségben vannak jelen a baktériumokban ahhoz, hogy befolyásolni tudják az adott genom evolúcióját.

A 11 bakteriális genomban található egyszerű szekvenciális ismétlődések eloszlásai eltérő mintázatot mutatnak, mely jellemző az adott (vagy több) törzsre. Bár a legtöbb tökéletes ismétlődés csak 12 nt hosszú, az ezeket határoló hosszabb nem tökéletes ismétlődések gyakran jeleznek múltbeli „kiterjesztés" (repeat expansion) és lebontás típusú evolúciós eseményeket. Egyéb jellemző jelenségek még a vizsgált törzsek nem tökéletes és a tökéletes ismétlődés tartalmában megfigyelt preferált SSR osztályok és variációk jelenléte, melyek világosan tükrözik a vizsgált genomokra irányuló különböző evolúciós nyomásokat. Az ismétlődés eloszlások alacsony hasonlósági értéke, különösen a *Chlamydiák*ban, azt mutatja, hogy az SSR preferenciák még közel rokon törzsekben is különbözhetnek, kialakítva egy új jelleget, mely a speciáció alatt változik. Eredményeink összhangban vannak azzal az elképzeléssel, hogy annak ellenére, hogy a mikroszatelliták kevésbé gyakoriak prokariótákban, mint eukariótákban, mégis befolyásolják a génevolúciót (Ellegren 2004).

A tökéletes és a nem tökéletes ismétlődések összehasonlításakor a legegyszerűbb feltételezés az, hogy mindegyik ugyanazt az összegzett (overall) genomi eloszlást követi, azaz csak a tökéletes ismétlődésekre kapott sokasági (abundance) adatok jól reprezentálják a nem tökéletes ismétlődésekre kapott adatokat. Ez igaz az általunk vizsgált genomok többségére, de meg kell jegyezni, hogy a *Chlamydia muridarum*ban az **agc** trinukleotid ismétlődéseknél tapasztalt tökéletes/nem tökéletes ismétlődés arány kiugróan magas a többi ismétlődés osztályhoz képest. Ez az eltérés a jelenlegi evolúciós változásokra utalhat, ha feltételezzük, hogy a tökéletes ismétlődésű „mag"-gal rendelkező nem tökéletes ismétlődések többsége tökéletes ismétlődésekként jelenhetett meg és a jelenlegi nem tökéletes ismétlődésű részek szétválasztó/szétszakító mutációk eredményei, így az eredeti SSR-ek még mindig azonosíthatók. Ezenfelül igen érdekes, hogy a trinukleotid ismétlődések eloszlását tekintve a *Chlamydia muridarum* valamiképpen hasonlít a vizsgált *E. coli* genomokhoz (3. kiegészítő táblázat), még akkor is, ha ez a gyenge hasonlóság csak az ismétlődések összhosszaiban mutatható ki. Úgy tűnik, hogy, az azonosított **agc** ismétlődések hajlamosak az kiterjedésre (expanzióra), a 13 érintett génből 9 tartalmaz mind tökéletes mind nem tökéletes **agc** ismétlődéseket (5. kiegészítő táblázat). Az érintett gének vagy a transzlációval kapcsolatosak (*infB* és *rplL*) vagy anyagcserével kapcsolatos fehérjéket kódolók (vagy még nincs

funkciójuk).

A mikroszatelliták aminosav ismétlések kódolása révén és a génexpressziót kontrolláló szekvenciákban való elhelyezkedés által hozzájárulhatnak a genetikai variációhoz (Kashi *és* King, 2006). A kódoló szekvenciák nagy hányada (és az ennek megfelelő relatív sok tri- és hexanukleotid ismétlődések) azt sugallják, hogy az előbbi mechanizmus gyakori a baktériumokban. Kimutattuk, hogy a preferált trinukleotid ismétlődések törzsenként különböznek és ezek a különbözőségek, mind a tökéletes, mind a nem tökéletes ismétlődések esetében megfigyelhetők (2. kiegészítő táblázat). Ez az SSR-eknek a bakteriális polimorfizmus kialakításában való aktív szerepét jelzi, egybevágóan a korábbi javaslatokkal (van Belkum *és mtsi.* 1998). Figyelemre méltó, hogy a tri- és hexanukleotid ismétlődések kódoló régióbeli jelenlétét a *Mycoplasma hyopneumoniae*ben nemrég közölték (Mrazek *és mtsi*. 2006), ami azt sugallja, hogy megállapításaink általánosabban is fontosak lehetnek a prokarióták körében.

Az ismételt szekvenciák okozta polimorfizmust korábban több *Chlamydia* genomban is vizsgálták (Rocha *és mtsi*. 2002). Eredményeink, melyeket egy másik megközelítéssel kaptunk, megerősítették ezeket a megállapításokat. A nagy SSR sűrűséggel rendelkező gének közül azok, amelyeket funkció szerint is besoroltak (a KEGG adatbázis osztályozási rendszere szerint), vagy alapvető sejtszintű folyamatokért felelősek, vagy polimorfikus membránfehérjéket kódolnak. Ilyen például az FtsK fehérje az az alapvető bakteriális ATPáz, mely a leánykromoszómák pontos szegregációjáért felelős sejtosztódás során (Iyer *és mtsi.* 2004). A többszörös ismétlődések jelenléte ebben a génben néhány rész (szegmens) gyors evolúcióját engedheti meg a megfelelő fehérjében az *E. coli*ban, de a *Chlamydiák*ban nem.

Megfigyelésünk, miszerint a nagy SSR tartalmú gének közül, azok amelyek a környezeti információ feldolgozással kapcsolatosak nagyobb tökéletes/nem tökéletes ismétlődés aránnyal rendelkeznek, mint a más funkciójú gének (6. táblázat) felveti azt a lehetőséget, hogy ezekben a kódoló szekvenciákban felgyorsult az evolúció. Ezt a feltevést támogatják a *Chlamydia* genomok polimorfikus külső membrán fehérjéi. A *Chlamydiák*ban a több törzsre jellemző viszonylag nagy ismétlődés tartalommal rendelkező gének szinte kizárólag polimorfikus membrán fehérjéket (*p*olymorphic *m*embrane *p*roteins, Pmps) kódolnak (5. kiegészítő táblázat). A Pmp-k a *Chlamydiales* rendre jellemző egyedi fehérjék közé tartoznak (Vandahl 2004; Gomes *és mtsi*. 2006). (Érdekességként jegyzem meg, hogy az összesen 59 *Chlamydiales*-specifikus fehérje listáját közzétevő cikk (Griffiths *és mtsi.* 2006) nem említi ezt a *Chlamydiales* rendre nézve unikális fehérjecsaládot). A Pmp gének által kódolt fehérjék gyakran hozhatók kapcsolatba a baktérium patogenitásával (Gomes *és mtsi.*

2004; Carlson *és mtsi*. 2005). Például Pmp_21 (más néven PmpD), a leghosszabb a *C. pneumoniae*-ban expresszált 21 Pmp közül, amely egy Gram negatív bakteriális fehérjéket exportáló fehérje családba, az ún. autotranszporterek közé tartozik. A PmpD N-terminálisa a *Chlamydia* külső membrán autotranszporter komponense, mely a bakteriális invázióban és a gazdában való gyulladás kialakulásában játszik fontos szerepet (Wehrl *és mtsi*. 2004). A *Pmp* gének megfigyelhető polimorfizmusának egyik forrása éppen mikroszatellita preferenciájuk lehet.

Érdekes továbbá az is, hogy az *E. coli*ban található TolA, amely szintén egy membrán(hoz) kötött fehérje, egy nagyon hosszú - bár- nem tökéletes ismétlődést tartalmaz, amely polimorfikus a vizsgált törzsekben. Ebben az esetben az ismétlődés jelenléte leginkább a kodon szelekciós nyomásnak tulajdonítható, bár nem zárható ki, hogy ez az ismétlődés képes volt mind kiterjedni/csökkenni ezen gének evolúciós története során. Más tandem ismétlődés- analíziseknek (de Castro *és mtsi*. 2006) megfelelően javasoljuk, hogy az SSR polimorfizmus hozzájárul a prokarióta membránfehérjék variabilitásához.

Annak érdekében, hogy megvizsgáljuk a horizontális géntranszfer (HGT) eseményeknek az ismétlődés eloszlásokra vonatkozó hatását, megnéztük az összes általunk detektált HGT érintett *Chlamydia* gént (Ortutay *és mtsi*. 2003), azaz vagy azokat a géneket, amelyek a különböző donor fajokból valamelyik *Chlamydia* törzsbe vagy többük közös ősébe HGT-vel „érkeztek", ahol ez alkalmazható/lehetséges volt (4. táblázat) és meghatároztuk ismétlődés-tartalmukat (az SSR adatok a 7. kiegészítő táblázatban találhatók). Azokban az esetekben, amikor valamelyik *Chlamydia* közös ős volt a befogadó, nem detektáltunk különbséget a nem tökéletes ismétlődésként is azonosított tökéletes és a tökéletes ismétlődésként is azonosított nem tökéletes trinukleotid ismétlődések között (azaz csak olyan ismétlődéseknél találtunk különbségeket, melyeknek nem volt más ismétlődésű „párja"). Más esetekben nem találtunk trinukleotid ismétlődéseket a HGT-vel megjelenő génekben, ezért a HGT nem lehet felelős azokért a genomok közötti különbségekért, melyeket a. 8. táblázatban tüntettünk fel. Ez alapján azt gondoljuk, hogy az itt közölt eredményeket és következtetéseket nem befolyásolták HGT események.

Sikerült kimutatni, hogy a vizsgált *E. coli* és *Chlamydia* törzsek karakterisztikus SSR eloszlást mutatnak, amely a tri- és a hexanukleotid ismétlődések relatív gyakoriságával jellemezhető. A tökéletes és a nem tökéletes SSR-ek szimultán analízise az ismétlődés osztályok eloszlásának különbözőségét tárta fel a *C. muridarum*ban a jelenlegi evolúciós folyamatok jeleként. Állításunkat, miszerint a mikroszatelliták még közel rokon törzsekben is hozzájárulnak a prokarióta genomi variabilitáshoz alátámasztja az SSR eloszlás mintázatok

tapasztalt különbözősége a rokon genomokban. A környezeti információ feldolgozással kapcsolatos gének, és különösen a *Chlamydia*-specifikus Pmp fehérjék az olyan jelölt szekvenciák, ahol az SSR-ek hozzájárulhatnak a vizsgált törzsek genetikai variációjához. A hipotetikus fehérjéket kódolók relevanciája további vizsgálatot igényel.

7. Felhasznált irodalom

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389-33402.

Angellotti, M.C., Bhuiyan, S.B., Chen, G. and Wan, X.F. (2007) CodonO: codon usage bias analysis within and across genomes. *Nucleic Acids Res*., **35**, Web Server issue W132-136.

Aussel, L., Barre, F.X., Aroyo, M., Stasiak, A., Stasiak, A,Z. And Sherratt, D. (2002) FtsK is a DNA motor protein that activates chromosome dimer resolution by switching the catalytic state of the XerC and XerD recombinases. *Cell,* **108**, 195-205.

Azad, R.K., Lawrence, J.G. (2007) Detecting laterally transferred genes: use of entropic clustering methods and genome position. *Nucleic Acids Res.*, **35**, 4629-4639.

Baron, S.(ed.) (1996) Medical Microbiology. 4th ed. University of Texas Medical Branch, Galveston, Texas (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mmed.TOC&depth=2) *letöltve 2008. február 14.*

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*., **27**, 573-580.

Bingen, E., Picard, B., Brahimi, N., Mathy, S., Desjardins, P., Elion, J. and Denamur, E. (1998) Phylogenetic analysis of Escherichia coli strains causing neonatal meningitis suggests horizontal gene transfer from a predominant pool of highly virulent B2 group strains., *J. Infect. Dis.,* **177**, 642–650.

Blattner, F.R., Plunkett G. 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) The complete genome sequence of Escherichia coli K-12. *Science,* **277**, 1453-1474.

Boussau, B., Karlberg, E.O., Frank, A.C., Legault, B.A. and Andersson, S.G. (2004) Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proc. Natl. Acad. Sci. U S A.*, **29**, 9722-9727.

Bush, R.M. and Everett,K.D.E. (2001) Molecular evolution of the Chlamydiaceae. *Int. J.* Syst. *and Evol. Microbiol*., **51**, 203-220.

Carlson, J.H., Porcella, S.F., McClarty, G., Caldwell, H.D. (2005) Comparative genomic

analysis of Chlamydia trachomatis oculotropic and genitotropic strains. *Infect Immun.* ,**73**, 6407-6418.

Carugo, O., Pongor, S. (2002) Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. *J. Mol. Biol.*, **315**, 887-898.

Cox, R.L., Kuo, C.-C., Grayston, J.T. and Campbell, L.A. (1988) Deoxyribonucleic acid relatedness of Chlamydia sp. strain TWAR to Chlamydia trachomatis and Chlamydia psittaci. *Int. J. Syst. Bacteriol.* **38**, 265-268.

Dalevi, D., Dubhashi, D. and Hermansson, M. (2006) Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures. *Bioinformatics*, **1**, 517-522.

Dean, T.J., Breman, J.G., Measham, A.R., Alleyne, G., Claeson, M., Evans, D.B., Jha, P., Mills, A. and Musgrove, P. (ed.) (2006) Disease Control Priorities in Developing Countries IBRD/The World Bank and Oxford University Press, Washington D.C. (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=dcp2.TOC&depth=2). *letöltve 2008. február 11.*

de Castro, L.A., Rodrigues, Pedroso T., Kuchiishi, S.S., Ramenzoni, M., Kich, J.D., Zaha, A., Henning Vainstein, M., Bunselmeyer Ferreira, H. (2006) Variable number of tandem aminoacid repeats in adhesion-related CDS products in Mycoplasma hyopneumoniae strains. *Vet. Microbiol.*, **116**, 258-269.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L. and McAdams, H.H. (2004) Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci. U S A*, **9**, 3480-3485.

Chen, Y., Timms, P., Chen, Y.P. (2007) CIDB: Chlamydia Interactive Database for cross-querying genomics, transcriptomics and proteomics data. Biomol Eng., 362doi:10.1016/j.

Clermont, O., Bonacorsi, S. and Bingen, E. (2000) Rapid and Simple Determination of the Escherichia coli Phylogenetic Group. *Appl. and Env. Microbiol.*, **66**, 4555–4558.

COGs: The Clusters of Orthologous Groups (COGs) Database: Phylogenetic Classification of Proteins from Complete Genomes Website. (ftp://ftp.ncbi.nih.gov/pub/COG).

Corsaro, D., Valassina, M. and Venditti, D. (2003) Increasing diversity within Chlamydiae. *Crit. Rev. Microbiol.*, **29**, 37-78.

Corsaro, D. and Venditti, D. (2004) Emerging chlamydial infections. *Crit. Rev. Microbiol.*, **30**, 75-106.

Donachie, W.D. (2002) FtsK: Maxwell's Demon? *Mol. Cell*, **9**, 206-207.

Eckert, K.A., Yan, G. (2000) Mutational analyses of dinucleotide and tetranucleotide microsatellites in *Escherichia coli*: influence of sequence on expansion mutagenesis. *Nucleic Acids Res*., **28**, 2831-2838.

Ellegren, H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet*., **5**, 435-445.

Errington J., Bath, J., Wu, L.J. (2001) DNA transport in bacteria. *Nat. Rev. Mol. Cell Biol.*, **2**, 538-545.

Everett, K.D.E., Bush, R.M. and Andersen, A.A. (1999) Emended description of the order Chlamydiales, proposal of Parachlamydiaceae fam. nov. and Simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family Chlamydiaceae, including a new genus and five new species, and standards for the identification of organisms. *Int. J. Syst. Bacteriol*., **49**, 415-440.

Everett, K.D. (2000a) Chlamydia and Chlamydiales: more than meets the eye. *Vet. Microbiol*., **75**, 109-126.

Everett, K.D. (2000b) Chlamydial taxonomy. At: The Fourth Meeting of the European Society for Chlamydia Research, Helsinki, Finland, August 20-23, (http://chlamydiae.com/docs/Chlamydiales/taxonomy.asp). *letöltve 2007. augusztus 5.*

Faguy, D.M., Doolittle, W.F. (1999) Lessons from the Aeropyrum pernixgenome. *Curr. Biol*., **2**, R883-886.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol*.,**17**, 368-376.

Feng, P, Weagant, S.D., Grant M.A. (2002a) Enumeration of Escherichia coli and the Coliform Bacteria. In: Bacteriological Analytical Manual (8th ed.), FDA/ Center for Safety and Applied Nutrition (http://www.cfsan.fda.gov/~ebam/bam-4.html, l*etöltve 2007. augusztus 5.*

Feng, P, Weagant, S.D., Grant M.A. (2002b) Diarrheagenic Escherichia coli. In: Bacteriological Analytical Manual (8th ed.), FDA/ Center for Safety and Applied Nutrition (http://www.cfsan.fda.gov/~ebam/bam-4a.html). *letöltve 2007. augusztus 5.*

Fischetti, V., Landau, G. Sellers, P. and Schmidt, J. (1993) Identifying periodic occurences of a template with application to protein structure. *Inf. Proc. Letters*, **45**, 11-18.

Fritsche, T.R., Horn, M., Wagner, M., Herwig, R.P., Schleifer, K.H. and Gautom, R.K. (2000) Phylogenetic diversity among geographically dispersed Chlamydiales endosymbionts recovered from clinical and environmental isolates of Acanthamoeba spp. *Appl. Environ. Microbiol*., **66**, 2613-2619.

Gáspári, Z., Ortutay, C., Tóth, G. (2007) Divergent microsatellite evolution in the human and chimpanzee lineages. *FEBS Lett.,* **581**, 2523-2526.

Gogarten, J.P. and Townsend, J.P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol*., **3**, 679-687.

Gomes, J.P., Bruno, W.J., Borrego, M.J. and Dean,D. (2004) Recombination of the genome of Chlamydia trachomatis involving the polymorphic membrane protein C gene relative to ompA and evidence for horizontal gene transfer. *J. Bacteriol.* **186**, 4295-4306.

Gomes, J.P., Nunes, A., Bruno, W.J., Borrego, M.J., Florindo, C., Dean, D. (2006) Polymorphisms in the nine polymorphic membrane proteins of Chlamydia trachomatis across all serovars: evidence for serovar D a recombination and correlation with tissue tropism. *J. Bacteriol*., **188**, 275-286.

Grayston, J.T., Kuo, C.-C., Campbell, L.A. and Wang, S.-P. (1989) Chlamydia pneumoniae sp. nov. for Chlamydia sp. strain TWAR. *Int. J. Syst. Bacteriol*., **39**, 88-90.

Greub, G. and Raoult, D. (2002) Parachlamydiaceae: potential emerging pathogens. *Emerg. Infect. Dis*., **8**, 625-630.

Griffiths, E. and Gupta, R.S. (2006) Lateral transfers of serine hydroxymethyltransferase (glyA) and UDP-N-acetylglucosamine enolpyruvyl transferase (murA) genes from free-living Actinobacteria to the parasitic chlamydiae. *J. Mol. Evol*., **63**, 283-296.

Griffiths, M., Ventresca, M.S. and Gupta, R.S. (2006) BLAST screening of chlamydial genomes to identify signature proteins that are unique for the Chlamydiales, Chlamydiaceae, Chlamydophila and Chlamydia groups of species. *BMC Genomics*, **7**, 14.

Gupta Lab's Bacterial Phylogeny Website (2008) http://www. bacterialphylogeny.com. (http://www.bacterialphylogeny.info/groupspecific/chlamydia/chlamydiae.html, http://www.bacterialphylogeny.info/groupspecific/chlamydia/chlamydiae_LGT.html). *letöltve 2008. február 5.*

Hale, T.L. (1991) Genetic basis of virulence in Shigella species. *Microbiol. Rev.,* **55**, 206–224.

Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating the human-ape splitting by a molecular clock of mitochondrial RNA. *J. Mol. Evol.,* **22**, 160-174.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M. and Shinagawa, H (2001) Complete genome sequence ofenterohemorrhagic

Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.,* **8**, 11-22.

Herzer, P.J., Inouye, S., Inouye, M. and Whittam, T.S. (1990) Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of Escherichia coli., *J. Bacteriol*., **172**, 6175–6181.

Hillis, D.M., Moritz, C. and Mable, B.K. (1996) Molecular Systematics, 2nd ed. Sinauer Associates, Sunderland, Massachusetts.

Horn, M., Wagner, M., Muller, K.D., Schmid, E.N., Fritsche, T.R., Schleifer, K.H. and Michel, R. (2000) Neochlamydia hartmannellae gen. nov., sp. nov. (Parachlamydiaceae), an endoparasite of the amoeba Hartmannella vermiformis. *Microbiology*, **146**, 1231-1239.

Horn, M. and Wagner, M. (2001) Evidence for additional genus-level diversity of Chlamydiales in the environment. *FEMS Microbiol. Lett*., **204**, 71-74.

Horn, M., Collingro, A., Schmitz-Esser, S., Beier, C.L., Purkhold, U., Fartmann, B., Brandt, P., Nyakatura, G.J., Droege, M., Frishman, D., Rattei, T., Mewes, H.W. and Wagner, M. (2004) Illuminating the evolutionary history of chlamydiae. *Science*, **304**, 728-730.

Hsia, R.C., Small, P.L. and Bavoil, P. M. (1993) Characterization of virulence genes of enteroinvasive Escherichia coli by TnphoA mutagenesis: identification of invX, a gene required for entry into HEp-2 cells. *J. Bacteriol*., **175**, 4817–4823.

Iyer, L.M., Makarova, K.S., Koonin, E.V. and Aravind, L. (2004) Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucleic Acids Res.,* **32**, 5260-5279.

Jukes, T.H. and Cantor, C.R. (1969) Evolution in protein molecules. In: *Mammalian Protein Metabolism.* (H.N.Munro, ed.) 21-123. Academic, New-York.

Jurka, J. and Pethiyagoda, C. (1995) Simple repetitive DNA sequences from primates: compilation and analysis. *J. Mol. Evol.,* **40**, 120-126.

Kahane, S., Greenberg, D., Friedman, M.G., Haikin, H. and Dagan, R. (1998) High prevalence of “Simkania Z” a novel Chlamydia-like bacterium,in infants with acute bronchiolitis. *J. Infect. Dis*., **177**, 1425-1429.

Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R.W., Olinger, L., Grimwood, J., Davis, R.W. and Stephens, R.S. (1999) Comparative genomes of Chlamydia pneumoniae and C. trachomatis. *Nature Genet*., **21**, 385-389.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354-357.

Kashi, Y. and King, D. (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends Genet*., **22**, 253-258.

Kassai-Jáger, E., Ortutay, C., Tóth, G., Vellai, T and Gáspári, Z. (2008) Distribution and evolution of short tandem repeats in closely related bacterial genomes. *Gene*, **410**, 18-25.

KEGG: Kyoto Encyclopedia of Genes and Genomes Website (http://www.genome.jp/kegg)

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol.Evol.,* **16**, 11-120.

Kimura, M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci.,* **78**, 454-458.

Kotloff, K.L., Winickoff, J.P., Ivanoff, B., Clemens, J.D., Swerdlow, D.L.,Sansonetti P.J., Adak, G.K.and Levine, M.M. (1999) Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. *Bull. WHO* **77**, 651–666.

Koski, L.B., Morton, R.A. and Golding, G.B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol*., **18**, 404-412.

Koonin, E.V. (2003) Horizontal gene transfer. The path to maturity. *Mol. Microbiol*., **50**, 725-727.

Lan, R., Alles, M.C., Donohoe, K., Martinez, M.B., Reeves, P.R. (2004) Molecular evolutionary relationships of enteroinvasive Escherichia coli and Shigella spp. *Infect. Immun.,* **72**, 5080–5088.

Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol*., **44**, 383-397.

Lawrence, J.G. and Ochman, H. (1998) Molecular archeology of the *Escherichia coli* genome. *Proc.Natl. Acad. Sci. USA*, **95**, 9413-9417.

Lawrence, J.G. and Ochman, H. (2002) Reconciling the many faces of lateral gene transfer. *Trends Microbiol*., **10**, 1-4.

Lawrence, J.G. and Hendrickson, H. (2003) Lateral gene transfer, when will adolescence end? *Mol. Microbiol.,* **50**, 739-749.

Leclercq, S., Rivals, E. and Jarne, P. (2007) Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics*, **8**, 125.

Lindstedt, B.A. (2005) Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. *Electrophoresis,* **26**, 2567-2582.

Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N., Weizenegger, M., Neumaier, J., Bachleitner, M. and Schleifer, K.H. (1998) Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis*, **19**, 554-568.

Massey, T.H., Aussel, L., Barre, F.X. and Sherratt, D.J. (2004) Asymmetric activation of Xer site-specific recombination by FtsK. *EMBO Rep.*, **5**, 399-404.

McNally, D., Fares, M.A. (2007) In silico identification of functional divergence between the multiple groEL gene paralogs in Chlamydiae. *BMC Evol. Biol.*, **7**, 81.

Metzgar, D., Thomas, E., Davis, C., Field, D. and Wills, C. (2001) The microsatellites of Escherichia coli: rapidly evolving repetitive DNAs in a non-pathogenic prokaryote. *Mol. Microbiol.*, **39**, 183-190.

Miklós, I. (2002) Statisztikus szekvencia illesztés. ELTE TTK Biológiai Doktori Iskola, Elméleti biológia és ökológiai Doktori Program, doktori értekezés

Mirkin, B.G., Fenner, T.I., Galperin, M.Y. and Koonin, E.V. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.*, **6**, 3:2.

Mrazek, J. (2006) Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. *Mol. Biol. Evol.,* **23**, 1370-1385.

Mrazek, J., Guo, X., Shah, A. (2007) Simple sequence repeats in prokaryotic genomes. *Proc. Natl. Acad. Sci. USA,* **104**, 8472-8477.

Nataro, J.P. and Kaper, J.B. (1998) Diarrheagenic *Escherichia coli*, *Clin. Microbiol. Rev.,* **11**, 142-201.

National Center for Biotechology Information (NCBI) Website (http://ncbi.nlm.nih.gov)

Noller, A.C., McEllistrem, M.C., Pacheco, A.G.F., Boxrud, D.J., Harrison, L.H. (2003) Multilocus Variable-Number Tandem Repeat Analysis distinguishes outbreak and sporadic Escherichia coli O157:H7 isolates. *J.Clin. Microbiol.*, **41**, 5389-5397.

Ochman, H., Lawrence, J.G. and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299-304.

Ochman, H. and Davalos, L.M. (2006) The nature and dynamics of bacterial genomes. *Science*, **311**, 1730-1733.

Ortutay, C., Gáspári, Z., Tóth, G., Jáger, E., Vida, G., Orosz, L. and Vellai, T. (2003) Speciation in *Chlamydia*: genomewide phylogenetic analyses identified a reliable set of acquired genes. *J. Mol. Evol.,* **57**, 672-680.

Ortutay Csaba Péter (2003) Genom evolúció Chlamydiákban – teljes genom szekvenciák filogenetikai vizsgálata. ELTE TTK Biológiai Doktori Iskola, Evolúciógenetika, evolúciós ökológia, konzervációbiológi Doktori Program, doktori értekezés

Palys, T., Nakamura, L.K. and Cohan, F.M. (1997) Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. *Int. J. Syst. Bacteriol*. **47**, 1145-1156.

Pál, C., Papp, B.and Lercher, M.J. (2005a) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer *Nat. Genet.,* **37**, 1372-1375.

Pál, C., Papp, B.and Lercher, M.J. (2005b) Horizontal gene transfer depends on gene content of the host. *Bioinformatics,* **21**, Suppl 2, ii222-ii223.

Pál, T., Al-Sweih, N.A., Herpay, M. and Chugh, T.D. (1997) Identification of enteroinvasive *Escherichia coli* and *Shigella* strains in pediatric patients by an IpaC-specific enzyme-linked immunosorbent assay. *J. Clin. Microbiol*., **35**, 1757-1760.

Pearson, W.R. (1990) Rapid and Sensitive Sequence Comparison with FASTP and FASTA. *Methods Enzymol.,* **183**, 63-98.

Perna, N.T., Plunkett, G.3rd, Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A. and Blattner, F.R. (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature*, **409**, 529-533.

Podani, J. (2003) A szárazföldi növények evolúciója és rendszertana, ELTE Eötvös Kiadó, Budapest.

Podell, S., Gaasterland, T. (2007) DarkHorse. a method for genome-wide prediction of horizontal gene transfer. *Genome Biol*., **8**, R16.

Read, T.D., Brunham, R., Shen,C., Gill, S.R., Heidelberg, J.F., White, O., Hickey, E.K., Peterson, J., Umayam, L.A., Utterback, T., Berry, K., Bass, S., Linher, K., Weidman, J., Khouri, H., Craven, B., Bowman, C., Dodson, R., Gwinn, M., Nelson, W., DeBoy, R., Kolonay, J., McClarty, G., Salzberg, S.L., Eisen, J. and Fraser, C.M. (2000) Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. *Nucleic Acids Res*., **28**, 1397-1406.

Read, T.D., Myers, G.S.,Brunham, R.C., Nelson, W.C., Paulsen, I.T., Heidelberg, J., Holtzapple, E., Khouri, H., Federova, N.B., Carty, H.A., Umayam, L.A., Haft, D.H., Peterson, J., Beanan, M.J., White, O., Salzberg, S.L., Hsia, R.C., McClarty, G., Rank, R.G., Bavoil, P.M. and Fraser, C.M. (2003) Genome sequence of Chlamydophila caviae (Chlamydia psittaci GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. *Nucleic Acids Res*., **31**, 2134-2147.

Reed, J.L., Famili, I., Thiele, I. and Palsson, B.O. (2006) Towards multidimensional genome annotation. *Nat. Rev. Genet*., **7**, 130-141.

Reid, S.D., Herbelin, C.J., Bumbaugh, A.C., Selander, R.K. and Whittam, T.S. (2000) Parallel evolution of virulence in pathogenic Escherichia coli. *Nature*, **406**, 64-67.

Rocha, E.P., Pradillon, O., Bui, H., Sayada, C and Denamur, E. (2002) A new family of highly variable proteins in the Chlamydophila pneumoniae genome. *Nucleic Acids Res*., **30**, 4351-60.

Rodrígez, F., Oliver, J.L., Marin, A. and Medina, J.R. (1990) The general stochastic model of nucleotide substitution *J. Theor. Biol.,* **142**, 485-501.

Rooney, A.P. (2003) Selection for highly biased amino acid frequency in the tolA cell envelope protein of Proteobacteria. *J. Mol. Evol.,* **5**, 731-736.

Salemi, M. and Vandamme, A.M. (ed.) (2003) The Phylogenetic Handbook: A Practical Approach to DNA and protein Phylogeny. Cambridge University Press, Cambridge, U.K.

Saunders, N.J., Boonmee, P., Peden, J.F. and Jarvis, S.A. (2005) Inter-species horizontal transfer resulting in core-genome and niche-adaptive variation within Helicobacter pylori. *BMC Genomics*, **27**, 9.

Schlotterer, C, Imhof, M., Wang, H., Nolte, V. and Harr, B. (2006) Low abundance of *Escherichia coli* microsatellites is associated with an extremely low mutation rate. *J. Evol. Biol*., **19**, 1671-1676.

Schmidt, H.A., Strimmer, K., Vingron, M. and von Haesler, A. (2000) TREE-PUZZLE Manual, Maximum likelihood analysis for nucleotide, amino acid, and two-state data Version 5.0.

Selander, R. K., Caugant, D.A. and Whittam, T. S. (1987) Genetic structure and variation in natural populations of Escherichia coli. In Neidhardt, F. C., Ingraham, J. L., Low, K.B., Magasanik, B., Schaechter, M. and Umbarger, H.E. (ed.), Escherichia coli and Salmonella typhimurium: cellular and molecular biology., American Society for Microbiology, Washington, D.C.

Sharp, P.M. and Li, W.H. (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflest selection for 'rare' codons. *Nucleic Acids Res*., **14**, 7737-7749.

Shirai, M., Hirakawa, H., Kimoto, M., Tabuchi, M., Kishi, F., Ouchi, K., Shiba, T., Ishii, K., Hattori, M., Kuhara, S. and Nakazawa, T. (2000) Comparison of whole genome sequences of Chlamydia pneumoniae J138 from Japan and CWL029 from USA. *Nucleic Acids Res.,* **28**, 2311-2314.

Sicheritz-Ponten, T, Andersson, S.G. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res*., **29**, 545-552.

Schleifer, K.-H. and Stackebrandt, E. (1983) Molecular systematics of prokaryotes. *Annu. Rev.Microbiol*., **37**, 143-187.

Snel, B., Bork, P. and Huynen, M.A. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.,* **12**, 17-25.

Stackebrandt, E., Goebel, B.M. (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species de®nition in bacteriology. *Int. J. Syst. Bacteriol*. **44**, 846-849.

Stanhope, M.J., Lupas, A., Italia, M.J., Koretke, K.K., Volker, C. and Brown, J.R. (2001): Phylogenetic analyses do not support horizontal gene transfer from bacteria to vertebrates. *Nature*, **411**, 940-944.

Stephens, R.S., Kalman, S., Lammel, C.J., Fan, J., Marathe, R., Aravind, R., Mitchell, W.P., Olinger, L., Tatusov, R.L., Zhao, Q., Koonin, E.V., Davis, R.W. (1998) Genome sequence of an obligate intracellular pathogen of humans, Chlamydia trachomatis. *Science,* **282**, 754-759.

Strimmer, K. and von Haesler, A. (1996) Quatret puzzling: a quartet maximum likelihood method for reconsructing tree topologies. *Mol. Biol.Evol*. **13**, 964-969.

Swofford, D.L. (2002) PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mithocondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512-526.

Taoka, M., Yamauchi, Y., Shinkawa, T., Kaji, H., Motohashi, W., Nakayama, H., Takahashi, N. and Isobe, T. (2004) Only a small subset of the horizontally transferred chromosomal genes in Escehrichia coli are translated into proteins. *Mol. Cell. Proteomics,* **3**, 780-787.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Thompson, D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penaltiesand weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Tóth, G., Gáspári, Z. and Jurka, J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967-981.

Tsirigos, A. and Rigoutsos, I. (2005) A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. *Nucleic Acids Res*., **8**, 3699-3707.

U.S. Centers for Disease Control and Prevention (U.S. CDC), U.S. Department of Health and Human Services (2005) Division of Bacterial and Mycotic Diseases: Foodborne illness US CDC Food Poisoning Guide Website (http://www.cdc.gov/ncidod/dbmd/diseaseinfo/foodborneinfections_g.htm). *letöltve 2007. szeptember 5.*

U.S. Centers for Disease Control and Prevention (U.S. CDC), U.S. Department of Health and Human Services (2007) Sexually transmitted diseases: Chlamydia (http://www.cdc.gov/std/chlamydia/default.htm). *letöltve 2007. szeptember 5.*

U.S. Food and Drug Administration Center for Food Safety and Applied Nutrition (U.S. CFSAN), U.S. Department of Health and Human Services (1992) Enterovirulent Escherichia Coli Group In: Foodborne Pathogenic Microorganisms and Natural Toxins Handbook (http://www.cfsan.fda.gov/~mow/chap13-16.html). *letöltve 2007. szeptember 5.*

van Belkum, A., Scherer, S., van Alphen, L., Verbrugh, H. (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev*., **62**, 275-293.

Vandahl, B.B.S., Birkelund, S. and Christiansen, G. (2004) Genome and proteome analysis of Chlamydia. *Proteomics*, **4**, 2831-2842.

von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M.,Foglierini, M., Jouffre, N., Huynen, M.A. and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*., **1**, 33 (Database issue): D433-437.

von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T.,Krüger, B., Snel, B. and Bork, P. (2007) STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*., **35**, Database issue: D358-362.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W.F., Surovcik, K., Meinicke, P. and Merkl, R. (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*, **16**, 142.

Wall, L., Christiansen, T. and Orwant, J. (2000) Programming Perl, Third Edition, O'Reilly, Sebastopol, USA.

Wehrl, W., Brinkmann, V., Jungblut, P.R., Meyer, T.F. and Szczepek, A.J. (2004) From the inside out--processing of the Chlamydial autotransporter PmpD and its role in bacterial adhesion and activation of human host cells. *Mol. Microbiol*., **51**, 319-334.

Welch, R.A., Burland, V., Plunkett, G.3rd, Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G.F., Rose, D.J., Zhou, S., Schwartz, D.C., Perna, N.T., Mobley, H.L., Donnenberg, M.S. and Blattner, F.R. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc. Natl. Acad. Sci.,* **99**, 17020-17024.

Whelan, S. and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691-699.

WHO (2007) Foodborne diseases (http://www.who.int/foodsafety/foodborne_disease/en/). *letöltve 2007. szeptember 5.*

Widenius, M. and Axmark, D. (2002) MySQL Reference Manual. O'Reilly Sebastopol, USA.

Woese, C.R. (2002) On the evolution of cells. *Proc. Natl. Acad. Sci. USA*, **99**, 8742-8747.

Xia, X. (2007) An improved implementation of Codon Adaptation Index. *Evol. Bioinformatics*, **3**, 53-58.

Yang, J., Nie, H., Chen, L., Zhang, X., Yang, F., Xu, X., Zhu, Y., Yu, J. and Jin, Q. (2007) Revisiting the Molecular Evolutionary History of Shigella spp. *J. Mol. Evol*., **64**, 71-79.

Zharkikh, A. (1994) Estimation of evolutionary distances between nucleotide sites. *J. Mol. Evol.,* **39**, 315-329.

Zhaxybayeva, O., Nesbø, C.L. and Doolittle, W.F. (2007) Systematic overestimation of gene gain through false diagnosis of gene absence. *Genome Biol*., **8**, 402.

8. Köszönetnyilvánítás

Szeretnék köszönetet mondani témavezetőmnek, **Dr. Vida Gábor**nak, aki mellettem állt és lehetővé tette, valamint támogatta munkámat melyet az ELTE Genetikai Tanszékén végeztem. Köszönöm **Dr. Vellai Tibor**nak, hogy segítette elindulásomat és utat mutatott, felhívta figyelmemet a horizontális géntranszfer jelenségére, valamint azt is, hogy kutatásunkat támogatta az OTKA F034933 és az FKFP 0041/2001 pályázatokból. Különösen hálás vagyok **Dr. Gáspári Zoltán**nak, aki mindvégig mellettem állt, hatalmas segítséget nyújtott a `PERL` programok írásával, valamint a mikroszatellita vizsgálatok és a CAI, CDT vizsgálatok végzése során. Ezentúl csoportvezetőjével, **Dr. Perczel András**sal együtt lehetővé tette, hogy a doktori ösztöndíj lejárta után az otthoni munka mellett az ELTE Szerveskémiai Tanszékén folytassam a mikroszatellita vizsgálatokat.

Köszönöm **Dr. Ortutay Csaba**nak a sokrétű segítséget, valamint hozzájárulását, hogy a HGT eredményekkel kapcsolatos közös publikációnk eredményei e dolgozatba is bekerülhessenek.

Köszönöm **Dr. Tóth Gábor**nak, hogy bevezetett a bioinformatika rejtelmeibe, a sok szakmai tanácsot, valamint azt, hogy lehetővé tette a hozzáférést a NIIF (*N*emzeti *I*nformációs *I*nfrastruktúra *F*ejlesztési *I*ntézet)-en keresztül a szuperszámítógépekhez, amit a *quartet puzzling* fák elkészítésénél vettünk igénybe.

Köszönet illeti családomat, akik mindvégig támogattak, lehetővé tették számomra a felkészülést, időt biztosítottak mind a kutatás számára, mind a dolgozat megírásához, amely kisfiam, **Levente** megszületése után igen nagy jelentőséggel bírt, neki is köszönöm, hogy elviselte „hiányomat". Nélkülük ez a munka és dolgozat nem készülhetett volna el.

9. A CD melléklet tartalma

Kiegészítő ábrák

1. kiegészítő ábra. A *groEL* génre készített `TREE-PUZZLE` kimeneti fájlok és a statisztikai analízis eredményfájlja [PDF formátumban]
A: `groEL_2.dat.puzzle`
B: `groEL_2.dat.tree`
C: `boostrap.log`

2. kiegészítő ábra. A trinukleotid ismétlődések megabázisonkénti teljes hossza [PDF formátumban]
A: A nem tökéletesként is azonosított tökéletes trinukleotid ismétlődések eloszlási mintázata az *E. coli* törzsekben
B: A tökéletesként is azonosított nem tökéletes trinukleotid ismétlődések eloszlási mintázata az *E. coli* törzsekben
C: A nem tökéletesként is azonosított tökéletes trinukleotid ismétlődések eloszlási mintázata a *Chlamydia* törzsekben
D: A tökéletesként is azonosított nem tökéletes trinukleotid ismétlődések eloszlási mintázata a *Chlamydia* törzsekben (megegyezik a dolgozaban szereplő 10. ábrával)

3. kiegészítő ábra. Az egyes trinukleotid ismétlődés osztályok genomi eloszlásának grafikus összefoglalása a vizsgált törzsekben [PDF formátumban]
A: *Escherichia coli* törzsek
B: *Chlamydia* törzsek

4. kiegészítő ábra. A *tolA* és az *ftsK* gén teljes nukleotid és fehérje szekvencia illesztése a 4 *E. coli* törzsben [PDF formátumban]
A: *tolA*
B: *ftsK*

Kiegészítő táblázatok

1. kiegészítő táblázat. A vizsgált baktériumok ismétlődés eloszlásának statisztikai adatai (a tökéletes és a nem tökéletes egyszerű szekvenciális ismétlődések gyakorisága az egyes baktérium törzsekben) [EXCEL fájl]

A: *Escherichia coli* CFT073, **B:** *Escherichia coli* K-12 MG1655, **C:** *Escherichia coli* O157:H7 EDL933, **D:** *Escherichia coli* O157:H7 Sakai **E:** *Chlamydia muridarum*, **F:** *Chlamydia trachomatis* D/UW-3/CX, **G:** *Chlamydophila caviae* GPIC, **H:** *Chlamydophila pneumoniae* AR39, I: *Chlamydophila pneumoniae* CWL029, **J:** *Chlamydophila pneumoniae* J138, **K:** *Chlamydophila pneumoniae* TW-183.

2. kiegészítő táblázat. Az azonosított trinukleotid ismétlődések által kódolt aminosav ismétlődésekre vonatkozó statisztika [EXCEL fájl]

A: Tökéletes ismétlődések
B: Nem tökéletes ismétlődések

3. kiegészítő táblázat. A nem tökéletes ismétlődésként is azonosított tökéletes ismétlődések és a tökéletes ismétlődésként is azonosított nem tökéletes ismétlődések eloszlási mintázatának összevetése az egyes törzsekben [EXCEL fájl]

A: Az ismétlődés eloszlások hasonlóságára vonatkozó számítások összefoglalása
B: A számítások részletei (kizárt osztályok, Khi-négyzet, stb.)

4. kiegészítő táblázat. A legnagyobb ismétlődés-tartalmú gének adatai [EXCEL fájl]

A: A baktériumonkénti 10 legnagyobb tökéletes ismétlődés-tartalmú gén
B: A baktériumonkénti 10 legnagyobb nem tökéletes ismétlődés-tartalmú gén

5. kiegészítő táblázat. A legnagyobb tökéletes ismétlődés-tartalmú gének funkcionális részletes adatai [EXCEL fájl]

A: A baktériumonkénti 10 legnagyobb tökéletes ismétlődés-tartalmú gén funkcionális adatai
B: A baktériumonkénti 10 legnagyobb nem tökéletes ismétlődés-tartalmú gén funkcionális adatai
C: Az `agc` trinukleotid ismétlődést tartalmazó gének a *Chlamydia muridarum* Nigg törzsben

6. kiegészítő táblázat. A polimorfikus memránfehérjéket kódoló gének [EXCEL fájl]

7. kiegészítő táblázat. A HGT érintett *Chlamydia* génekben azonosított SSR-ek adatai [EXCEL fájl]

A: A HGT érintett gének adatai **B:** CPn0486, **C:** dppF, D: groEL, **E:** pfrA, **F:** yceC, **G:** ychB

10. Kivonat

Az egyre több prokarióta genomi szekvencia megismerésével az utóbbi években lehetővé vált ezen szervezetek evolúciójának részletes és átfogó kutatása. A prokarióta genomikában eleinte a horizontális géntranszfer (HGT) került az érdeklődés középpontjába olyan folyamatként, amely jelentős szerepet játszhat egy mikróba génkészletének kialakulásában, így a genomszerkezet evolúciójában. A napjainkban egységesen elfogadott filogenetikai analízis - amellyel eldönthető egy adott génről, hogy HGT eseménnyel került-e a genomba, illetve tájékoztatást nyújt a HGT irányáról is - tűnik a legobjektívebb és legpontosabb megközelítésnek. Ezek az evolúciós események közel rokon baktériumtörzsek teljes genom szekvenciáinak analízisével detektálhatók. A *Chlamydiales* rend 5 genomjának valamennyi fehérjét kódoló nyílt leolvasási keret (ORF) szekvenciáját vizsgáltuk hasonlósági kereséssel és filogenetikai analízissel. Sikerült megbízhatóan kimutatni egy teljes szekvenciakészletet, melynek tagjai a *Chlamydia* vonal divergálódása óta jelentek meg. Tudomásunk szerint ez az első szisztematikus filogenetikai alapú megközelítés, mely prokariótákban a szerzett gének megbízható kimutatását tűzte ki célul. Bár a *Chlamydiák* a magasabb rendű eukarióták obligát intracelluláris parazitái, s ez által feltehetőleg jobban elzártak a HGT-től, mint a szabadon élő fajok, eredményeink megmutatják, hogy diverzifikációjukban szerepet kapott idegen szekvenciák genomjaikba való felvétele is.

A genomi evolúció másik érdekes aspektusa, az egyszerű szekvenciális ismétlődések (SSR-ek, mikroszatelliták) vizsgálata. Az egyre több prokarióta genomszekvencia egy új típusú részletes összehasonlító vizsgálatot tett lehetővé. Tekintettel arra, hogy a mikroszatellitákat főként eukariótákban vizsgálták, egy genomszintű mikroszatellita-eloszlás vizsgálatot végeztünk 7 *Chlamydia* és 4 *Escherichia coli* genomban. Az általános trendek és a specifikus gének összehasonlítása ezekben a közeli rokon baktériumokban megmutatja, hogy a mikroszatelliták fontos szerepet játszanak a prokarióta genomok és számos géncsalád evolúciójában. Az egyes törzsek jellegzetesen eltérő mikroszatellita-eloszlást mutatnak, egyes, tipikusan polimorfnak ismert gének minden vizsgált genomban tartalmaznak ilyen típusú ismétlődéseket. A tökéletes és a nem tökéletes ismétlődések összehasonlítása fényt derít a mikroszatellita evolúció részleteire ezekben a genomokban. Eredményeink rámutatnak, hogy ezek a szekvenciák nem csupán eukariótákban, hanem prokarióta szervezetekben is jelentősen hozzájárulnak a genomok evolúciójához. A mikroszatelliták tehát az egész élővilágban hasonló módon formálják az élőlények genetikai anyagát.

11. Abstract

A possible way of detailed and comprehesive examination has been opened lately with coming to know of the increasing prokaryotic genomial sequences. It was *H*orizontal *G*ene *T*ransfer (HGT) that has become the observed of all observers in prokaryotic genomics, having been a process which may play an important role in the development of certain microbial gene sets, thus in the evolution of the genome structures. The present-day consensus is that phylogenetic analysis of individual genes is still the most objective and accurate approach for determining the occurrence and directionality of HGT. These evolutionary events can be detected with total genome sequence analysis of closely related bacterial species. We analyzed all protein-encoding ORFs of five genomes of *Chlamydiales*, performing similarity searches and phylogenetic analysis. We managed to identify a reliable set of sequences that have arisen via HGT since the divergence of the *Chlamydia* lineage. According to our knowledge, this is the first systematic phylogenetic inference-based attempt to establish a reliable set of acquired genes in a bacterial genome. Although *Chlamydia* are obligate intracellular parasites of higher eukaryotes, and thus suspected to be isolated from HGT more than the free-living species, our results show that their diversification has involved the introduction of foreign sequences into their genome.

Examination of *S*imple *S*equence *R*epeats (SSRs, microsatellites) is another interesting aspect of genome evolution. A new possible way of detailed and comprehensive examination has been opened lately with the growing number of prokaryotic genome sequences. Since microsatellites were examined mainly in eukaryotes earlier, we performed a genome wide analysis of microsatellite - distribution in 7 *Chlamydial* and 4 *Escherichia coli* genomes. Comparison of overall trends and specific genes in these closely related bacteria shows that microsatellites play an important role in the evolution of prokaryotic genomes and several gene families. The individual strains show characteristic microsatellite distribution. Genes known to be typically polymorphic contain repeats in all examined genomes. Comparison of the corresponding perfect and imperfect repeats also sheds light on the details of the evolution of microsatellites in these genomes. Our results point out that these sequences significantly contribute to the evolution of the genomes not only in eukaryotes but also in prokaryotes. Microsatellites form the genetic material of living beings similarly in the whole world of life.